SENTIMENTALISM AND MORAL MOTIVATION

by

NILE SEDGWICK

(Under the Direction of Sarah Wright)

ABSTRACT

Sentimentalism is the view that moral judgment has something essentially to do with response, feeling, or sentiment. On a rational version of this view, judging an action or practice to be morally wrong to do is equivalent to believing that guilt would be a fitting response. Moral judgment thus becomes the imposition of standards about how to feel. This way of looking at things affords a nonreductive account of moral thought. Moral claims are not construed as themselves sentiments are dispositions to feel, but judgments about what is fitting to feel. This provides the resources to make sense of the essential contestabilty of moral concepts as well as a explanation for why motivation tends to coincide with moral belief. The following essay is a defense of a rational sentimentalism and a demonstration of how the analysis could be used to explain why we are typically motivated to do what we believe we ought.

INDEX WORDS: Sentimentalism, Emotion, Metaethics, Moral Motivation, Fittingness

SENTIMENTALISM AND MORAL MOTIVATION

by

NILE SEDGWICK

B.A., University of Georgia, 1999

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

SENTIMENTALISM AND MORAL MOTIVATION

by

NILE SEDGWICK

Major Professor:	Sarah Wright
Committee:	Charles Cross
	Beth Preston

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

DEDICATION

For my father

Acknowledgements

Kevin Zaragoza helped me a lot in developing the project early on. Sarah Wright, who advised me most along the way, generously offered her expertise and patience. Charles Cross and Beth Preston graciously gave their time and effort serving on the committee. Finally, to friends and family who kept me company during the long while it took to finish, thanks.

Table of Contents

Page

Acknowledgements.....v

List of Tables.....viii

Introduction.....1

A. Background.....1

B. Morality and Blame.....8

C. Guilt and Morality.....11

E. Preview.....12

Chapter

1 Rational Sentimentalism and Fittingness.....19

1.1 Sentimentalism and the Metaethical Landscape.....19

1.2 Why a Rational Sentimentalism?.....33

1.3 The Response Dependency Thesis and Fittingness.....45

1.4 The Wrong Kinds of Reasons.....53

2 Judgment and Response.....80

2.1 Essential Contestibility and Univocity.....80

2.2 Emotions and Judgments.....86

2.3 Guilt and Representation.....103

3 Moral Motivation.....117

3.1 The Problem.....117

3.2 Internalism and Motivation.....123

3.3 Externalism and Motivation.....137

3.4 Rationalism and Motivation.....150

3.5 Sentimentalism and Motivation.....172

References.....184

List of Tables

Page

Table 1: Ontological Distinctions.....31

Table 2: Sentimentalist Distinctions.....32

Introduction

A. Background

The following is an essay on moral language. It is an evaluation and defense of an analysis of moral claims. It is not an assessment of those acts or practices that are right or wrong. It is not an evaluation of those reasons we have for calling something 'right' or 'wrong,' nor does it deal with the question of whether we are ever justified in taking a stand on a moral issue. It is not, in short, a discussion of morality *per se* but of what we mean by calling some act or practice 'right' or 'wrong.' What are we doing when we use moral terms? What exactly are we saying when we proclaim that something is 'wrong' to do? This essay offers an answer to these questions.

The upshot has implications for a further problem in moral philosophy. Moral conviction bears in an interesting relation to what we are motivated to do. It would be odd if one were to assert a moral belief but admit that she simply did not care not at all about that. We would expect this person to condemn those who disagreed, or at least to think unfavorably about them. If she herself were in a position to consider doing what she thought was wrong, we would expect her to be motivated, at least to some degree, not to do it. It seems strange to imagine her saying something like "I know that this is wrong, but so what? I really don't care if I or anyone else does it that matter." Most would be inclined to think that this person is not really sincere in her

belief. At the very least, the oddity of this kind of attitude suggests that motivation and moral belief typically go together. The view that I defend here offers an explanation for why this is so.

I work with a view that descends roughly from Hume. Hume thought that morality has something essentially to do with how we feel. This he took to be phenomenologically obvious:

> vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflexion into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action...Here is a matter of fact; but 'tis the object of feeling, not of reason.... It lies in yourself, not in the object. (1739/1978:468-469)

What this means depends upon many of the other things Hume thought and said, none of which I'm going to look into here. But this particular insight is still useful. Morality does seem to have something to do with our feelings. It's hard to deny that moral dispute is often very emotional. Sometimes it's so emotional that nothing ever gets resolved. Taking sides on a moral issue is often a source of both interpersonal and intrapersonal conflict and dilemma, and of marking off lines between friends and enemies. It is natural to associate feelings of guilt towards ourselves at having done something we believed was wrong, or anger at others for violating norms that require us to act in certain ways. Whatever is meant by finding vice in a 'sentiment,' feelings have a very close relationship to moral belief.

Hume is now considered the founder of 'sentimentalism.' Sentimentalism is a view about moral judgment that treats motivating attitudes as somehow essential. The view has recently found a resurgence in the works of Gibbard (1990), Blackburn (1998), Wiggins (1987), and

especially D'Arms and Jacobson (2000), just to name a few. These thinkers authors have recently offered what they call a 'rational' perspective on the position. Rational sentimentalism is concerned with a certain kind of normative constraint on attitudes towards value objects, where the content of this relation is one of fittingness or appropriateness. An object has value, on this view, just in case some attitude is a fitting response to it. It has been articulated in the following response-dependency thesis (RDT):

> RDT: to think that X has some evaluative property Φ is to think it appropriate to feel F in response to X (D'Arms and Jacobson 2000a: 729).

What makes RDT a sentimentalism is the content of F. The attitudes in question belong to the set of motivating attitudes directed to or about a specific range of value features (emotions). Fear, for example, is about something whose features constitute its status as threatening or dangerous. In the present vernacular, fear is 'characteristically concerned' with these value properties. Nondiscernable attitudes, on the other hand, comprise the most talked about mental states in metaethical discussions. These are things like beliefs and desires, whose content is not limited to any particular set of value features, but could be about anything. There is nothing, for instance, about desiring something that tells us anything about the value the object other than the fact that one desires it. The hallmark of a nondiscerning attitude is that its content doesn't necessarily give expression to its grounds. The object of a discerning attitude, like fear, does. Fear is characteristically concerned with dangerous things. Its content specifies the range of value objects that would justify feeling it.

In this essay, I want to extend RDT to the moral case. Morality falls under the general structure of value, and has often been contrasted with things that have value. Much has been made, for instance, of the distinction between the 'right' and the 'good' (cf. Rawls 1971). Where the 'right' denotes something of what we may or ought to do, the 'good' has something to do with the way world should be, or what is desirable. Though it doesn't seem that things could be right or wrong, they could be good or bad. Considerations such as these have often compelled thinkers to treat moral issues separately from value, or the good. And I want to do just the opposite.

The term 'value,' however, has a wide range of meanings. It refers to a wide variety of both moral and nonmoral notions. 'Values issues,' for example, seem to have some connection with what we take to be right and wrong. Someone with strong values might be said to hold esteem for personal reliance and self sufficiency, adhere to a traditional conception of the family, etc. But of course many other nonmoral things are commonly said to have value, such as nature, art, scientific inquiry and the like. Close ties to one's friends and family is a valuable thing, though perhaps not a moral requirement. We also attribute value to attitudes and deeds that, though not morally required, seem good to have. Generosity is valuable. Loyalty, trustworthiness, having skill in a trade, respecting the environment—all these things are valuable.

Morality is a species of value that has many different guises. So-called 'thick' moral concepts seem to require their own unique treatment. There is something intrinsically different between the 'repugnant' and the 'shameful.' Some acts are 'disgusting' in a moral sense, some are 'honorable,' some 'disgraceful' and others 'austere.' There are countless thick designations. 'Thin'

moral concepts, by contrast, are only two, RIGHT and WRONG. A thick moral concept such as 'contemptible' refers to the property to which contempt is fitting. Thus, thick concepts specify in the description of that property the response that is appropriate. Thin concepts, on the other hand, do not. Concepts RIGHT and WRONG do not describe attitudes at all.

Nevertheless, there are attitudes that are fitting responses to actions that have the properties that RIGHT and WRONG refer to. According to one prominent view (Gibbard 1990), an action is wrong if it would be appropriate to feel guilty for doing it and to resent another for doing it. This essay is essentially an endorsement for this view. Moreover, I limit the focus of the discussion solely to self-directed moral claims in order to distill the connection between moral belief and motivation. If Sally judges that Harry's wrongdoing merits resentment on his part, this says little about her motivation not to Φ. But if Sally judges that Φ-ing would be wrong for her to do, the fact that guilt would be fitting—so she believes—speaks directly to her resistance to Φ-ing. The prospect for feeling guilt is unpleasant. It is this negative emotional outcome that we seek to avoid when considering whether to engage in what we take to be a moral transgression, so I argue.

On a moral version of RDT, the evaluative property Φ, will be the property WRONG, while F will be the concept GUILT. 'Wrong' and 'guilt' bear out the fittingness relation between response and moral property in self-directed moral claims. While guilt is the sentiment fit to moral transgression, the concept WRONG picks out the property that guilt is a fitting response to. On a moral version of RDT, when one believes that Φ-ing is wrong to do then one believes that guilt would be a fitting response to Φ-ing, or:

RDT_m: to think that Φ-ing is wrong is to think it appropriate to feel guilt in response to Φ-ing.

Again, the analysis defended here is narrow. It is just one side of a larger analysis that includes the other-directed attitude of resentment. Thus, everything that pertains to the defense offered here goes *mutatis mutandis* for the other-directed form. I focus only on the self-directed side to illuminate the source of concern we ourselves have for doing what we believe is right. A good deal of energy has been spent on explanations for this phenomenon and I want to offer my own take on what an explanation might be. This is not to say of course that by developing the other side of the analysis we don't get something informative about self-directed motivation. Indeed, the attitudes of guilt and resentment likely go hand in hand. Consider the words of Gibbard on this subject:

> Take guilt and resentment: if one person resents an action of another and the other does not feel a corresponding guilt, we may expect trouble. Guilt makes possible the acknowledgment of wrong, and such modes of reconciliation as restitution, compensation, apology, and forgiveness. One's chances of damaging conflict are reduced, then, if one feels guilty when guilt and its normal accompaniments are demanded by others, and if one demands guilt and its normal accompaniments only when others are prepared to feel guilty. Hence it tends to be advantageous for an individual to coordinate his guilt with the resentment of others and his resentment with the guilt of others. (1990:67-68)

The effort in this essay is aimed at detailing the side of this dynamic that involves the tendency to make reparations for one's action. Again, the focus here is limited in order to address the problem of moral motivation, which is a self-directed concern to do what one believes one ought. But this concern likely exists in tandem with other-directed moral concerns. Thus, though self-directed moral claims and their corresponding motivations can be assessed independently of their other-directed counterparts, it (the assessment of self-directed moral claims) ultimately comprises only a part of a larger picture.

Further, as I mentioned above, there are many kinds of moral judgments we make that do not appear in the thin guise of 'wrong.' We regularly judge actions to be shameful in a moral sense, or disgusting or contemptible and a host of other things. I don't address these kinds of judgments here, but RDT is tailor made for them. On a response-dependent analysis of these thick assessments, to judge something shameful, for example, would be to think that shame is appropriate or fitting. One might appropriately be ashamed at being overweight, or lazy or self-absorbed, or whatever. Thick evaluative claims are not necessarily moral claims. And since I want to apply an analysis of moral claims to the problem of moral motivation, I restrict the discussion to RDT_m.[1]

Despite the limited range of this analysis, the scope captures a very relevant range of moral assessment. Judgments about whether an act or practice would be wrong to do are widespread and comprise among the most controversial topics. We argue about the thin moral status of war, abortion, gay rights, and capital punishment often in attempts to establish or undermine a legal basis for them. In our personal lives we encounter dilemmas that involve theft, adultery, lying, and much worse. The concept WRONG thus has a very wide application.

1 See revision below.

Accordingly, guilt is a prevalent and powerful response whose conditions of fittingness is something many of us are invested in knowing.

B. Morality and Blame

Wrongful action can be committed by those who cannot be blamed for committing them. Suppose, for example, that Tommy is very depressed. Say he's encountered some sort of trauma recently. Adding to Tommy's troubles is the fact that his friend has just landed in the hospital with a terminal illness. Say that Tommy is morally obliged, and takes himself to be obliged, to visit his dying friend. But he doesn't go because he's depressed. In this case, it seems that Tommy acts wrongly (or is wrong not to act) yet blamelessly. His extenuating circumstances exonerate him from criticism we would otherwise rightly aim in his direction. In this case, it is not appropriate to feel guilty for doing what Tommy believes is wrong.

The sentimentalist might try to resist this conclusion. It's plausible to insist, at least, that Tommy hasn't done anything wrong because he isn't responsible for his inattentiveness. 'It's not your fault,' we might tell him, 'you haven't done anything wrong.' But even if we disregard competing intuitions about this case, there are other cases that make this line of reasoning implausible. Imagine Tommy in a different set of circumstances. Suppose that he is responsible for the lives of millions and that inattentiveness to their concerns would dramatically affect their quality of life. Despite his extenuating circumstances, it seems blame would be appropriate. Indeed, the distinction is most painfully clear when one is forced to wrongdoing against their will. If, instead of being depressed, Tommy were somehow forced to ruin the lives of millions—

say by threat to his family—it would be even less likely that we would exonerate him for doing something wrong. Though more trivial cases of wrongdoing may not clearly bring out the distinction between wrongdoing and blameworthiness, severe wrongful action does.[2]

We might wonder why we need an account of believing wrong separate from blame. But these two concepts have distinctively different applications. Blame is a backward-looking notion. When we blame someone we criticize them for what they've done. Considerations about acts that we suspect are wrong to do, on the other hand, are often forward-looking. We wonder whether Φ-ing would be wrong to do and why. Of course, one might wonder and reason about whether they have done something wrong. But by far the most important application of the concept is to what we are to do. Morality is a practical discourse at its core. The morally conscientious agent tries to figure out which alternatives are right to do and which ones are wrong, and then tries to choose the ones that are not wrong. She's concerned with questions that pertain to what's to be done, not generally about what has been done.

Examples of blameless wrongdoing encourage the thought that the analysis is better equipped to handle claims about blame than about wrongdoing. When we say that guilt is an appropriate response to Φ-ing we recommend it (guilt), apparently as a measure of censure. That is, in most cases this is what we do when we say that Φ-ing is wrong to do. If one cannot be blamed for the wrong one does, then there is little in the way of censure that would seem appropriate. Wrongdoing becomes merely unfortunate in this kind of case—something that happened as a result of extenuating circumstances whose absence would have affected the outcome.

2 For a forceful statement of this distinction, see Arendt (1977).

On the other hand, there is clearly a connection between moral wrongness and blameworthiness. When we say that Φ-ing is wrong to do we typically mean that one would be blameworthy for doing it. It may turn out that he would not be blameworthy if his circumstances warrant excusing him. But, for the most part, blame and wrongdoing go hand in hand. When the morally conscientious agent deliberates about what to do, he himself thinks in these terms. If, for example, Tommy were to ask himself whether to visit his friend in the hospital, we could imagine him saying 'yes, of course. I should visit her because she's my friend, after all. I should feel guilty for not doing so.' Indeed, this would be the case even if Tommy were very depressed. Though we would be likely to excuse his inattentiveness, he likely would not.[3]

RDT_m effectively analyzes moral claims as claims about blame. Since blameless wrongdoing does occur, however, it needs to be revised. Claims about moral wrongness can be translated into claims about blameworthiness in the following way. We can say that one acts wrongly if they would be *prima facie* blameworthy for Φ-ing, where the phrase '*prima facie*' means before the circumstances surrounding the agent's motivations have been taken into account (Gibbard 1990). Consider finally $RDT1_m$:

> $RDT1_m$: To think that Φ-ing is wrong is to think it *prima facie* appropriate to feel guilt in response to Φ-ing.

The distinction between blameworthiness and wrongdoing suggests that a restriction to the analysis is needed to accommodate cases where RDT_m breaks down.

3 At the very least, it would be odd of him if he did.

C. Guilt and Morality

Another worry about the account is that guilt is not a specifically moral emotion. D'Arms and Jacobson have pressed this intuition. Consider the following example:

> Mother has grown older, and grown mentally ill. She makes increasingly exigent demands on the family. Her illness is degenerative. She always feared being "put away"; you know she wants to stay at home, but you have real doubts about your ability to care for her. And you also see the effects of the tension, and finally resentment on your family. Your spouse, who has been patient and helpful throughout, is beginning to show the strain. The children are restless. On any plausible normative picture, there are ample reasons for the conclusion you've been avoiding. In the end, you decide to put her in a nursing home. You're convinced this is the right thing to do, though you know you'll feel guilty for doing it. But does your guilt make sense? If you think it does, and yet that your action was not wrong, there are problems for the analysis. (1994:742-743)

The authors think that it is fitting or appropriate to feel guilt in this situation and yet that one not believe that they have done anything wrong.[4] Are they correct? The example is controversial. On the one hand, it shows the intuitive plausibility behind the idea that guilt is not a uniquely moral emotion. One may think, quite reasonably, that guilt is fitting in this case. But, on the other hand, one may also disagree. Clearly, it makes sense that you feel guilty. Putting your mother in a nursing home against her will would seem to elicit guilt. That is to say, there

4 This contributes to an explanation for why D'Arms and Jacobson reject RDT_m (1994).

are explanatory reasons for feeling guilt here. There is not, however, a justifying reason for feeling it. It's understandable that you feel guilt but not rational in the sense of being appropriate or fitting. The example is thus best thought of as a borderline case where guilt seems fitting without any moral concern for what's to be done. Nonetheless, intuitions are strong on both sides. While it's certainly reasonable to think that guilt is fitting in this case, it is also reasonable to disagree. Thus, the example is not strong enough to falsify the analysis.[5]

It is natural to construe guilt as a moral emotion. There is not a single psychologist to my knowledge who presents a nonmoral version of guilt.[6] Further, there is evidence linking guilt to a conceptual hold on right and wrong.[7] But the most important point is that intuitively the concept is not problematic. Guilt is generally felt in response to doing what one believes is wrong. It appears to require a belief system built around right and wrong such that we are justified in saying that guilt is an appropriate response to perceived wrongdoings. Borderline cases notwithstanding, it is reasonable to assume that guilt is a moral emotion.

D. Preview

Much has to be done to explain what it means to say that guilt would be fitting. Indeed, the notion itself is difficult to get a hold of. For, without restricting fittingness to a certain range of reasons it's ambiguous. It might be fitting, for instance, if by feeling guilty one were able to contribute to the good of herself or others. Say that you buy someone a gift for their birthday but

5 And the authors themselves readily admit this: “We can accommodate normative disagreement here; we won't dig in our heels against all who deny that guilt is appropriate” (D'Arms and Jacobson 1994:744).

6 Not everyone takes a stand on the issue, of course. But for those who speculate about emotional contents, guilt is essentially tied to wrongdoing.

7 I briefly discuss this point later in Chapter Two, Section **2.3**

it's not what they wanted. Unfortunately, this person is deeply offended. He seeks compensation and apology. If by feeling guilty for what you've done increases the likelihood of meeting his demand, guilt might be fitting. But this is not the right kind of reason to feel guilty. Instead, guilt is justified in the right way only if felt in response to something that's morally wrong. Guilt, we might say, is characteristically concerned with moral transgression. This suggests that the right kinds of reasons for feeling guilt are those that speak to the relationship between the object of one's response and the response itself. It is these reasons that speak to an emotion's fittingness.

But this also reveals a more pressing problem with the analysis. If the right kinds of reasons for feeling guilt depends in some way on its propositional content, and if the content of guilt involves the concept WRONG, then analyzing assessments about wrongdoing in terms of assessments about guilt just takes us back to where we started. $RDT1_m$ in other words, appears to be circular. Judgments involving the concept WRONG are analyzed into judgments involving the concept GUILT which itself contains the concept WRONG. However, like all feelings, guilt is not a simple propositional attitude. Rather, it has two components. Alongside its characteristic concern we can identify a feeling-aspect. Guilt feels a certain way; it's not just directed towards a certain type of property. By describing the emotion according to its affect we get an analysis of moral belief that's informative. Moral belief can be analyzed into belief about the fittingness of guilt in a way that exploits the conceptual relationship between the concept WRONG and guilt's characteristic concern while bringing in its affective quality.

The upside is considerable. First, we gain a way of thinking about moral belief that makes good sense out of a salient feature of moral conviction and dispute. We can explain the

fact that two people can disagree on a moral issue when wielding different ideas of what it means to do something wrong. Consider a likely debate between consequentialist and deontologist over whether the killing of an innocent person would be morally wrong. The consequentialist would say that such an action would not be wrong if the outcome were beneficial. The deontologist would disagree, arguing that such an act violates a moral rule not to kill innocents. These two disagree on this issue because they have different standards of WRONG that they bring to the table. But it's not as if they're not actually disagreeing about something. If we can reference the concept guilt by its feeling aspect, then $RDT1_m$ offers us a way of understanding what's going on here. What these two are really arguing about can be cast in terms of whether 'this feeling' (whatever it's like to feel guilt) is an appropriate response to killing an innocent person in these circumstances.

Furthermore, the structure of guilt makes it an ideal candidate for explaining why it is we are motivated to resist doing what we believe is wrong. The feeling side of it has a tie to our motivations. The negative prospects of feeling guilt should provide at least some motivation to resist whatever it is that would elicit it. The conceptual side to it, on the other hand, should render an explanation for why it is that our moral beliefs and feeling of guilt typically coincide. They share the same content, after all, and would likely be elicited by a commitment to the same facts.

So here's what's to come. In Chapter One I focus on where this particular brand of sentimentalism falls on the metaethical landscape. As a view that incorporates our responses it is essentially subjective. And since it aims solely to analyze moral judgment, it remains neutral on some controversial issues in the metaethical literature today. As well, I point out some salient

differences between the view I endorse and other sentimentalist incarnations. In the latter part of the chapter I turn to the task of explicating the notion of fittingness. As I mentioned above, emotions can be justified in various ways. But a response is only fitting if it is justified in the right way. The challenge of separating out the right kinds of reasons for having a response from the wrong kinds of reasons has been called 'the wrong kinds of reasons problem.' I offer a solution to this problem.

In Chapter Two I turn to an examination of the sentiment involved in $RDT1_m$, guilt. One of the promising upsides of the analysis is its ability to explain how moral dispute can remain connected when disparate parties are using different concepts of WRONG. In order to succeed in this task, it is necessary that the concept GUILT be formulable in a way that does not incorporate the same concept we want to analyze. An inability to do so not only threatens the univocity of moral dispute but renders the analysis circular. I argue that guilt has two components. On the one hand is its propositional content. Guilt's content does involve the concept WRONG. But on the other side is the feeling affect. While guilt can be construed as an attitude characteristically concerned with moral transgression, it also has a distinctive feel. And by referencing its feeling-aspect when making judgments about the fittingness of guilt we get a way out of the vicious circularity that threatens the analysis. We can also make sense of how two people with different ideas of what counts as 'wrong' can nonetheless be moored in moral dispute. Following Prinz (2004a), I endorse a theory of emotion that makes sense of these two sides. According to this particular view, the distinctive feeling of an emotion has the function of carrying information about value-laden features of the world. Guilt can involve the concept WRONG, then, in the sense

of having the specific function of detecting what we take to be instances of moral transgression while remaining type-identifiable by its affective quality.

In Chapter Three I apply the analysis to the problem of moral motivation. Philosophers have long tried to explain the connection between what we believe we ought to do and what we are motivated to do. It doesn't exist in all people. Some are not motivated to do what they think is right because they they have some psychological difficulty that impedes what would otherwise be a concern to do what they think they ought. Others simply don't care about what's right or fail to appreciate reasons they have to act. But for many of us there is a connection between what we believe we ought to do and what we are motivated to do. The question is, why?

Michael Smith (1994) has argued that difficulty finding an explanation for this phenomenon suggests a fundamental 'moral problem.' The relationship between moral belief and motivation is usually cast in terms of a relationship between two different kinds of mental states: belief and desire. But belief and desire have different directions of fit. For example, if I believe that there is beer in the fridge and discover that there is not my belief is likely to change. If, on the other hand, I desire that there is beer in the fridge and discover that there is not then my desire will likely not change. Thus, while belief is counterfactually dependent on evidence that *not-p*, desire is not. The fact that they typically go together then, is mysterious. I offer a solution to the moral problem that in effect makes it disappear. Instead of construing motivation in terms of desire, I propose that we understand it in terms of emotion. Emotions, such as guilt, have the function of carrying information about the world. This gives motivation and belief the same world-to-mind correspondence and affords an explanation of moral motivation via the

relationship between what our moral beliefs purport to express and what our feelings are designed to do.

The story very roughly goes like this. By judging an action as wrong we take ourselves to have reason to feel guilty. If we are disposed to feel guilty for doing what we believe is wrong, the prospects of feeling guilt involves taking on an emotional cost. In so far as we don't want to incur this cost, we will not be motivated, at least to some degree, to do what we believe is wrong. Thus, as commitments to the fittingness of guilt, moral claims have implications for the motivations of people who are not listless or fail to appreciate what reasons they have to act. That is to say, it explains the coincidence of motivation and moral belief in people who are 'normal' in this sense.

Other accounts of moral motivation fail in this task for one reason or another. I evaluate two recent approaches to solve the problem of moral motivation and show why they fall short. On one view, the connection between motivation and moral belief has nothing to do with the content of our moral beliefs but rather by contingent psychological facts about us. This account, I argue, doesn't allow for the sensitivity our motivations have to reasons we have to Φ. Our motivations are sensitive to grounds in ways that imply that the content of our moral claims really do matter. On another view, moral belief is understood as a commitment to motivations we are rationally required to have. But this account doesn't do either, since it is not always rational to be motivated to do what we think is morally required.

There is much to recommend a rational sentimentalism about moral judgment. By thinking of moral claims in terms of responses that we should have (they're fitting), we are able

to explain the interesting connection between behavior and moral conviction. And this, if anything, is what a theory of moral judgment should do.

Chapter One

Rational Sentimentalism and Fittingness

1.1 Sentimentalism and the Metaethical Landscape

Rational sentimentalism is a distinctive metaethical view. It is subjective in its attempt to analyze value judgment in terms of response. Yet exactly how it does this distinguishes it not only from objective views about morality—which purport to leave response out altogether—but also from other sentimentalist views. Value claims do not just involve response. Rather, they are claims about response. They are claims that some response would be fitting.

One of the most significant features of sentimentalism is what it does not explicitly state. To judge that something has value is simply to say that some response is fitting. The claim does not express a commitment to the idea that fittingness relations are facts of any sort. Nor does it entail that we project fittingness attitudes onto the objects of our judgment. Rather, sentimentalism is neutral on the realism debate.

Accordingly, the view can be formulated in one of two ways, either as a cognitivism or a noncognitivism about moral judgment. Cognitivism about moral judgment has it that they are fact-stating (Brink 1989, Sturgeon 1985, Mackie 1977). Value judgments are descriptive of value facts, facts that they may or may not correctly refer to. On a noncognitivist reading, moral

judgments are not fact-stating but are expressions or recommendations of *pro* or *con* attitudes (Blackburn 1993, Gibbard 1990).

Cognitivist versions of sentimentalism are called 'sensibility theories' (Wiggins 1987, McDowell 1987). There are two core claims of sensibility theory. First, values are perceived. Sensibility theorists believe that value properties exist and are perceived, but that value has something intrinsically to do with the valuer. Consider an analogy to color vision. Colors inhere in objects as real properties in the world. Yet, color is just the property of an object as it appears to normal people under normal circumstances and in this way depend on the viewer. Sensibility theorists diverge from the color analogy slightly by insisting that there is some normative space between what we perceive and what we are entitled to believe—as McDowell claims, "a virtue (say) is conceived to be not merely such as to elicit the appropriate 'attitude' (as color is merely such as to cause the appropriate experiences), but rather such as to *merit* it (1985, original italics)." Still, their view is a cognitivism about value, and lends itself to the idea that values are secondary properties.

Second, sensibility theorists are cognitivists about the attitudes they use to analyze value judgment. They espouse what might be called a 'no priority' view about emotion and evaluative judgment. On this view, one cannot understand shameful without shame, wrongness without guilt, etc. But the reverse is also true, shame is just that response concerned with the shameful, guilt with wrongdoing, etc. Though this is admittedly circular, we gain some new understanding of evaluative terms by formulating them in terms of how we are to respond to doing what we believe is valuable. Admiration is not just the belief that one has some strength or accomplishment, its the peculiar quality of that feeling (whatever that is), as well as the

disposition to praise those whom you admire, to nominate them for a citizenship award, or whatever. The affective qualities and behavioral tendencies of these responses provide us with something more than what we might expect from the concept WRONG.[8]

Noncognitivist varieties of sentimentalism are commonly called 'projectivisms.' Notable proponents are Blackburn (1993) and Gibbard (1990). In contrast to sensibility theory, projectivism also has two salient features. First, it's an antirealism about value. There are no value facts or properties on this view. Rather, we project our attitudes onto the world.[9] Second, projectivism gives priority to our responses. It analyzes value judgment in terms of some attitude that is appropriate or fitting, where that attitude does not involve the judgment we are trying to analyze. Projectivists thus offer a noncircular analysis of value judgment that's based on a noncognitivist reading of our responses.

Sentimentalism also differs from other kinds of cognitivism and noncognitivism. It diverges from any account whose implications for moral judgment entirely ignores response. Such is the case, for example, with a naturalist moral realism. The core claim of this view is that moral facts obtain independently of our evidence for them, that there is nothing about us or how we perceive the world that could make a difference as to whether something is right or wrong to do (Sturgeon 1986, Brink 1989). Relativism would therefore be false. It could not be the case that different parties who had all of the relevant natural facts of the matter draw different conclusions about whether Φ-ing is right to do and all be correct. For the naturalist moral realist,

8 It's well worth asking whether we should expect anything more from an analysis of value judgment. C.f. Wiggins when he says, "when we consider or not *x* is good or right or beautiful, there is no appeal to anything that is more fundamental than actually possible human sentiments...But what use (I shall be asked) is such a formulation? My answer is that, by tracing out such a circle, the subjectivist hopes to elucidate the concept of value by displaying it in its actual involvement with the sentiments. One would not, according to him, have sufficiently elucidated what value is *without* that detour (1987:229).

9 Projectivism descends from Hume, who argues against the idea that values are percieved. See 1975:291-292.

morality is not a study directed at our responses, but with the external, natural world (Brink (1989:18,n.5).[10] Further, the view implies cognitivism, since our beliefs about morality will be true or false depending on what the moral facts are.

Noncognitivist non-sentimentalisms also permeate the landscape. Consider Hare's prescriptivism (1952). Hare interprets moral claims, such as the claim that it is right for A to Φ, into prescriptions of the form 'let A Φ.' These expressions are noncognitive in the sense that they do not have a truth value. But they are not expressions of attitudes that the agent has. Rather, they are recommendations or allowances for particular actions.

The distinctions between these theories turn on ontological commitments. They differ with respect to intuitions about whether there are any value facts of the matter, and what our claims about them actually are. Sensibility theory and projectivism are competing views over what value claims are and whether there are any facts that correspond to them. One has it that such claims are fact-stating and sometimes true, the other thinks they are projections of attitudes and never true nor false. The fact that sentimentalism cuts across this distinction indicates that it is not a thesis about ontology. Rather, sentimentalism is a semantic view about value claims. It's neutrality on the realist debate is due to the difference between a semantic approach to understanding value language and an ontological view as to what those claims are.

As a semantical claim, sentimentalism is cognitive. Indeed, each view discussed so far is cognitive in this sense. They all respect the intuition that normative judgment appears in a property ascriptive discourse that purports to state the facts. Moral judgment may actually be projections of attitudes or prescriptions for action, but they seem to be reports of moral facts. This much should be accommodated by any analysis of moral claims.

10 I push the difference between response-dependent theories and narutalist moral realism further in section **2.1.**

Importantly, the semantic cognitivist commitment is not necessarily to value or moral facts of a non-natural sort. We generally take the natural properties of an object to constitute its value. Consider the words of Ewing on this subject:

> the reason why it is proper to admire anything must be constituted by the qualities which make the object of admiration good, but it does not follow that the thought that it is good must, if the admiration is to be justifiable, intervene between the perception of the factual qualities admired and the feeling of admiration. (1947:158)

And later he says,

> The ground [for admiration] lies not in...goodness, but in the concrete, factual characteristics of what we pronounce good. Certain characteristics are such that the fitting response to what possesses them is a pro attitude, and that is all there is to it. (1947:172)

Ewing's insight has more recently been called by Scanlon a 'buck-passing account:'

> Being good, or valuable, is not a property that itself provides a reason to respond to a thing in certain ways. Rather, to be good or valuable is to have other properties that constitute such reasons. Since the claim that some property constitutes a reason is a normative claim, this account also takes goodness and value to be non-natural properties,

> namely the purely formal, higher-order properties of having some lower-order properties that provide reasons of the relevant kind...It is not goodness or value itself that provides reasons but rather other properties that do so. For this reason I call it a buck-passing account. (1998:97)

On a buck passing account, the relevant moral property of Φ-ing is just the property of Φ-ing of having some natural properties that constitute reason for thinking that Φ-ing is wrong. 'Wrongness' is the higher-order property of Φ-ing whose lower-order, natural properties justify thinking that Φ-ing is wrong. For example, say that a few of the neighborhood kids are pouring gasoline on a cat and are about to light it on fire.[11] The lower-order properties of this act, among them the setting of a cat on fire, constitute reason for thinking that doing such a thing is wrong. Setting the cat on fire does not also have some lower-order, non-natural property 'wrongness.' Rather, it has only natural features which justify thinking that committing this hideous act is wrong. The normative buck is passed from the higher-order 'wrongness' to the lower-order natural features in virtue of which it is wrong.

Updating the reading of admiration, *pace* Ewing, we might say that this emotion is characteristically concerned with a strength or achievement on one's part. But instead of referring to some non-natural property of achievement, we may refer to the natural properties of Φ-ing in virtue of which it displays strength or achievement—where the terms 'strength' and 'achievement' are read as evaluative terms we are entitled to use in a description of Φ-ing. According to a buck-passing account, we can use them to fill out the content of admiration if we

11 This unfortunate example is taken from Harman (1977).

wish, but our attitude is fit to those natural properties in virtue of which we can describe Φ-ing as exhibiting one's strength or achievement.[12]

This way of describing the matter is in contrast to the account given by Moore (1903). Moore argues that the value of something is an irreducible, non-natural property. Value properties are *sui generis*, on his view, distinct from the natural properties of an object. The natural features of an object cannot themselves give us reason to respond, since we can always ask whether these natural features are 'good,' and thus whether we ought to respond to them in some way. Goodness is thus itself is a reason-giving property. On a Moorean account, we should respond negatively to acts of cruelty, not because of any natural features that such acts have, but because they have the property of wrongness. The value that they have is itself reason to disapprove of them.

For what it's worth, the idea that there are any facts or properties other than natural ones is now widely rejected.[13] Many think that when it comes to ascertaining value, it seems sufficient just to refer to those natural properties that make a thing good or valuable or right or wrong. The fact that Φ-ing would cause needless suffering, for example, seems perfectly well to justify not doing it, or feeling guilty for doing it, in the sense that this natural property provides a complete explanation of the reasons we have for responding in these ways. It is not clear, on the other hand, what further work could be done by a special, reason providing, non-natural property

12 Of course, this doesn't mean that one's admiration couldn't be due to the thought that Φ-ing is indicative of one's strength. There's a difference between what the characteristic concern of an attitude is, and what one's attitude might be elicited by on some occasion. Suppose that one admires another simply because he believes the other to be successful. Say he doesn't know why she's successful but just believes that she is. In this case, one admires another not because of properties she has that justify thinking she is successful, but just because one believes her to have these properties (whatever they are). But the 'because' here is merely causal. One's belief that another is successful here elicits admiration, but does not justify it. Further, one's admiration may or may not be fitting in this case—it all depends on whether she really is successful.

13 Naturalism is embraced by realists and antirealists alike. See Railton (1986), Brink (1989), Gibbard (1990), Blackburn (1993).

of wrongness. Further, many different actions can be said to be right or wrong and the grounds for these judgments vary widely. But there does not seem to be one simple, unanalyzable property common to all actions that are wrong to do.[14]

The fact-stating appearance of moral claims motivates cognitivists. Consider Brink:

> As many have observed, moral discourse is typically declarative or assertive in form. We say things like 'The government's tax plan is unfair', 'Waldo is just', 'It would be wrong to work for that cause'...Our moral judgments not only have fact-stating property-ascribing *form*; they have cognitivist *content* as well...it is often claimed that one should not be held responsible for actions one could not have known were wrong, that goodness deserves reward, that turpitude of a crime should determine the severity of the punishment...In making such moral judgments, we (or at least those who make them) certainly seem to presuppose the existence of moral facts and properties and the possibility of moral knowledge. (1989:25-26)

Brink thinks that the natural starting point for a metaethic is to assume that there are moral facts, that our claims about them can misrepresent them, and give up on this commitment only if there are very good reasons to think it isn't true.

Noncognitivists begin by denying the existence of moral facts and the corresponding commitment to moral language. This puts a burden on them to explain how it is that our moral claims appear to presuppose the existence of moral facts. Thus, Blackburn develops a

14 I often speak of an object having 'evaluative properties.' This is meant in a sense neutral to the realist debate. I don't intend to take sides on the issue of whether moral properties are non-natural, natural, or even whether there are any moral properties at all.

'quasirealist' view about moral judgment (1987). Moral claims are projections of attitudes about how to respond, and so do not represent moral facts. Yet they retain all the trappings of cognitivism and realism. They can function as property ascriptions well enough and express the objectivity associated with moral conviction. But much work has to be done to explain how noncognitive attitudes can be embedded in a purportedly fact-stating discourse and how we could be entitled to say that something is ever really right or wrong to do.[15]

In any event, the commitment that projectivists have to the semantical aspects of moral language separate them from nihilism, the view that nothing is ever right or wrong to do. Nihilism requires an antirealism about morality, but is not entailed by it.[16] Consider Blackburn on the subject:

> [the antirealist] affirms all that could ever properly be meant by saying that there are real obligations...it is just that the explanation of why there are obligations and the rest is not quite that of untutored common sense. It deserves to be called anti-realist because it avoids the view that when we moralize we respond to, and describe, an independent aspect of reality...[M]athematics provides a useful model for understanding this. There are anti-realist views of what we are doing when we practice arithmetic. But they need not and should not lead to anyone wondering whether 7 + 5 is 'really' 12, for that would be an expression of first-order doubt that would not be a consequence of the second-order theory. (1973:157)

15 This task may be a considerable disadvantage to the position. See Dreier (1990).
16 Nietzsche appears to occupy the position of a moral nihilist. See his *On the Genology of Morals* (1887/1967).

The validity of Blackburn's claim, if true, owes to the distinction he makes between first-order and second-order moral discourse. First-order moral discourse comprises the semantical realm. It's the kind of language we use in our everyday normative talk, such as when I say that 'stealing from the cookie jar is wrong.' It's the sphere of discourse in which we argue about what things are right or wrong to do, and make claims about what we believe is wrong and why. Second-order moral discourse is directed towards reflections on ontology. It is the theory we use to analyze or describe what our moral claims actually mean. On an ontological level, my claim that 'stealing from the cookie jar is wrong,' could mean a number of things. I might actually be describing a property of stealing from the cookie jar. Or perhaps I merely attempt to describe a property of pilfering a cookie but fail; there may not actually be any such property. Another alternative is that I'm not describing anything at all but rather expressing some *con* attitude about stealing from the cookie jar. Any of these three interpretations of what I mean are examples of second-order theorizing. If Blackburn is right and whatever second-order theory we adopt says nothing necessarily about how moral terms must be used in first-order discourse, then we can believe that moral terms are simply expressions of attitudes we have, that there are no moral facts or properties in the natural world, and still use moral terms as if both of these beliefs were false. We can separate out our second-order moral views from our first order ones in the same way that the mathematician who believes that there are no mathematical facts doesn't have to balk at the idea that 7 + 5 is 'really' 12.

Thus, when Blackburn says that the antirealist "affirms all that could ever properly be meant by saying that there are real obligations" he means that the antirealist can successfully account for how our moral language on the first order of business seems to be exactly like that a

realist about morality would use.[17] If I say that 'stealing from the cookie jar is wrong,' I mean, on the first order of discourse, that it's wrong, full stop. It's wrong for me to do and it would be wrong for anyone to do in relevantly similar circumstances. It's wrong even if I believed that it wasn't, etc. I can even say that it's 'true' that stealing from the cookie jar is wrong. But on the second order of discourse I need not assent to any of these claims. I might be a noncognitivist about morality and take myself to merely expressing my disapproval of stealing from the cookie jar. I might believe that there is no moral fact of the matter that stealing from the cookie jar is wrong to do. But I need not be somehow insincere in my commitment to the obligation I take myself in having not to steal, or give up on any of the intuitions that a realist enjoys accommodating. My second order views about morality need not infect my first order moral convictions.

Again, however, there has to be some grounds that entitle one to make the first order claim. Projectivists need to explain how our moral claims can be, though not true or false, correct in some sense. Cognitivist realists don't have a difficult time of this, since appeal to the moral facts would readily explain why Φ-ing would be 'really' wrong to do. Nihilism, meanwhile, denies that any moral claim is ever true (or correct).

Nihilism has some convergence with another prominent metaethical view: error theory. Error theory is a cognitivism about moral judgment conjoined with an antirealism about morality (Mackie 1977). The error theorist holds that our moral claims are fact-stating but systematically fail. Error theory is thus a form of nihilism, but not all nihilists are error theorists. When the nihilist makes her case against the projectivist, she does not disagree on the cognitive status of moral judgment. Rather, she attacks projectivist grounds for believing that Φ-ing is wrong to do.

17 Thus the name of his view, 'quasi-realism.'

These grounds don't amount to an absence of any moral fact of the matter, since they both are antirealists. The nihilist meets the projectivist on her own terms and holds that the justification given for any moral claim falls short.

Rational sentimentalism cuts across these metaethical distinctions. By leaving the ontological question alone, it is consistent with cognitivism and noncognitivism about normative judgment, realism and antirealism about value facts and the corresponding denial of the truth or correctness of any value claim. A nihilist rational sentimentalist, for instance, would analyze value claims into claims about the fittingness of some emotion, but then hold that this emotion is never fitting. She might do this by arguing that the fittingness claim tries to pick out the facts but fails, or because the projectivist program fails. The view avoids disputes over the ontological status of value claims and the corresponding question over whether they are ever justified.

More importantly, given the controversial nature of these divisions, sentimentalism doesn't beg the question against cognitivist or noncognitivist, realist or antirealist. Indeed, the fact that much suspicion has crept into the debate among sentimentalist varieties of these positions encourages the thought that neutrality on the debate is a substantial advantage of the view. D'Arms and Jacobson, for instance, have rightly challenged the importance of focusing on this theme. As they put it "the [rational] sentimentalists all agree that values are not absolute or intrinsic properties, but rather are subjective and anthropocentric" and that, moreover, noncognitivists now see little reason to "forgo property talk and are even willing to speak of the truth of evaluative claims" (2000a:730).[18] The debate has become a little obscured by the willingness and ability of cognitivists and noncognitivists alike to accommodate the same

18 The leading noncognitive proponents of neosentimentalism are Blackburn and Gibbard.

intuitions. Where noncognitivists claim to be able to explain the apparent property-ascriptive quality of our moral claims, cognitivists are willing to accept that moral properties have no reality independent of us. The differences between the two thus appear pretty thin, and perhaps is still getting thinner. But what they have in common is their endorsement of RDT. Analyzing moral judgment in terms of judgments about reasons for having a certain response does not beg the question against either a cognitivist or noncognitivist theory of value discourse. Cognitivists and noncognitivists can argue over whether fittingness judgments are fact-stating. But whoever turns out to be correct has no bearing on whether RDT is true.

A diagram of the distinctions made so far might be helpful:

Table 1, Ontological Distinctions:

	Cognitivism	**Noncognitivism**
Realism	Naturalist Moral Realism Sensibility Theory	
Antirealism	Error Theory (Nihilism)	Prescriptivism Projectivism

In addition to differences between realists, antirealists, cognitivists and noncognitivists, there are important differences between rational sentimentalism and non-rational sentimentalisms. Expressivists, for example, such as Ayer (1936) or Stevenson (1937), assert that moral claims are merely expressions of *pro* or *con* attitudes. On their view, when I say that

Φ-ing is right to do, what I mean is something like 'yea Φ-ing.' I'm just giving expression to how I regard the act of Φ-ing. Since my attitudes can't be true or false, expressivism is not a cognitivism about moral judgment. Neither is it a realism about morality. Yet as a sentimentalism it is markedly different from rational sentimentalism. Expressivism analyzes moral claims into the having of an attitude, while a rational sentimentalism analyzes moral claims into claims about the fittingness of an attitude. The relevant difference between them is the normative space that the latter allows between claim and response.

Another kind of non-rational sentimentalism is a dispositionalism about value. According to this kind of view, value claims are just those responses one would have under certain conditions. Some have it that these conditions are what circumscribe normal people under normal circumstances (Prinz 2004b, Dreier 1990). Others prefer to idealize these conditions, and speak only of what someone with certain refined qualities that she exercises in a cool hour would decide (Hume 1965). Again, rational sentimentalism insists on retaining the normative force of value claims in its analysis. Instead of interpreting value claims in terms of how one is disposed to respond, a rational version of sentimentalism has it that such claims are about what responses are appropriate. To represent these further distinctions, consider:

Table 2, Sentimentalist Distinctions:

Rational Sentimentalism	**Non-Rational Sentimentalism**
Sensibility Theory	Expressivism
Projectivism	Dispositionalism

These distinctions are basic, informative, and should be helpful. In the next section, I argue that rational sentimentalism has an advantage over non-rational versions. A weakness of these accounts is the absence of any space for reasons to play a role in the having of an attitude. Rational sentimentalism allows for this space, and avoids taking on the burden of explaining how having or being disposed to have a response to Φ-ing can accommodate the normative force of the claim that Φ-ing has value.

1.2 Why a Rational Sentimentalism?

Why accept a rational sentimentalism?[19] What motivates the need for it? It's useful to compare the rational version of sentimentalism to non-rational sentimentalisms. This illuminates the historical development of the sentimentalist tradition and why it has arrived at a rational version of the view.[20]

A good starting place is the position normally called 'simple subjectivism.'[21] Simple subjectivism has it that moral statements are reports of one's approval or disapproval about some action or practice. The claim that abortion is morally wrong would be analyzed as the self-report of one's disapproval of the practice of abortion. This view appeals to the dispositions of the agent to reveal the meaning of moral claims. Accordingly, it has no problem explaining the motivational efficacy of believing wrong, since to believe wrong is just to be appropriately

19 From here on I refer to the view simply as 'sentimentalism,' unless otherwise appropriate.

20 I don't discuss advantages of a rational sentimentalism over nonsentimentalist views. But there are two ways in which this essay can be read as making this point. In section **2.1**, I discuss how a rational sentimentalism can accommodate some core intuitions about moral discourse. These are difficult to explain on a theory that does not make use of our responses in the analysis of moral judgment. Also, in section **3.3**, I develop a sentimentalist interpretation of the phenomenon of moral motivation. This too is difficult to explain on a nonresponse-dependent analyasis.

21 As discussed in Ayer (1936)

motivated. Believing wrong is just disapproving, i.e. to be disposed to think, act and feel in certain ways in certain circumstances.

The problem with simple subjectivism is that has a hard time explaining how people might actually disagree about a moral issue. If one person says that 'Φ-ing is wrong,' when all she means is 'I disapprove of Φ-ing,' and another person says that 'Φ-ing is right,' when all she means is 'I approve of Φ-ing,' neither is saying anything that could conflict with the claim of the other. Neither could be wrong in their descriptions unless they were deceiving themselves or lying about their own states of mind.

This problem is avoided by analyzing moral statements into expressions of approval or disapproval, instead of self-descriptive reports.[22] This idea, otherwise known as 'emotivism,' is natural, I think, since we often use normative words to express our feelings (such as with interjections). We might also describe these expressions as signs of dispositions the agent has, and like the simple subjectivist, link value claims to motivations. But emotivism is a bit more complicated than just this. Consider two groups of fans each rooting for their own team at a football game. When the home team makes a good play, supporters of the home team may express their approval, while supporters of the visiting team express their disapproval. Here we have expression of opposing attitudes without dispute (D'Arms and Jacobson 2000a: 725).[23]

What the expressivist really needs to capture is the way moral language is used with the intention to persuade that others feel the same as the one who uses it. Stevenson thus focuses on what he calls the 'dynamic' function of language (1937:21). This term is meant to describe the way language can be used to create moods (as in poetry), or to incite people to action (as in

22 Proponents of this view are Ayer (1936) and Stevenson (1937).

23 Emotivism can be construed as an expressivism about feelings or emotions one has or dispositions to feel, act, and think in certain ways. The latter, I think, is more plausible

oratory). A moral claim thus has to be understood as something like 'this is right, do so as well.'[24] It has to involve some sort of outward-looking prescription that others feel the same as oneself. Only then can the emotivist explain how opponents in moral dispute can be said to attempt to influence each other's attitudes. The one who thinks that "X is wrong" is saying something like "X boo," while the one who claims that "X is right" is saying something like "X yea." But each is not just expressing one's own attitudes. Each is trying to get the other to feel the same as she does; they disagree in attitude.

But there is a decisive problem with this kind of account. It is not able to explain those cases in which evaluative judgment and feeling come apart. Sometimes our emotions are at odds with what we deem valuable, and sometimes we deem valuable without feeling anything at all. Consider two examples given by the D'Arms and Jacobson:

> We can imagine a self-aware anorexic who finds herself feeling ashamed of how fat she is, against her better judgment. She is ashamed of her body, but she does not judge it to be shameful. Now imagine a disaffected sloth who has grown lazy, sedentary, and obese. Suppose, though that he nevertheless judges it desirable to have a healthy body weight and shameful how far he deviates from this ideal. Yet he cannot be bothered to eat well or exercise. After a period of being ashamed of himself, he has become disaffected and can no longer even muster shame at his weight, though he still judges it to be shameful. (2000a:726)

24 Stevenson points out, however, that even this is not wholly accurate, since the imperative in this statement makes an appeal to the conscious efforts of the hearer, and ethical language is more subtle than this. This is why he ultimately concludes that it is impossible to define 'good' in terms of favourable interest if its meaning is not to be distorted. See Stevenson (1937:26).

The self-aware anorexic and the disaffected sloth indicate that it is neither necessary nor sufficient to be in the relevant emotional state in order to make the corresponding value judgment. And the fact that emotivism is dispositional doesn't help, since the case of the disaffected sloth shows that one need not be even prone to feel the relevant emotion in order to make the corresponding value judgment.

The emotivist might respond that these two unfortunate people don't really token moral judgments. As a noncognitivist view about morality, emotivism simply has it that value statements are neither true nor false, but expressions of one's own attitudes about a particular issue. It is therefore impossible, by definition, to make a moral claim absent the relevant disposition.[25] However, this maneuver begs the question against one who offers these cases as counterexamples. The question of whether one can believe that something would be shameful without being disposed to feel shame cannot be answered just by stating one's position on this matter. The emotivist needs an independent reason for thinking that these cases are not ones of genuine value judgment—he can't just appeal to his own analysis of moral judgment to assess them.

One possibility is that these people don't really token moral belief but report what other people believe to be shameful. Suppose, for example, that a temp-agency manager asserts the belief that sex discrimination in the workplace is morally wrong, but regularly harasses his female employees. It's likely he's just paying lip service to a value everyone generally accepts. He knows that he must appear to accept it too, lest he be regarded a scoundrel. This guy doesn't really believe that women are due totally equal treatment in the workplace, he just knows to say he believes such a thing so that his character remains in good public standing. This is the

25 This is why Mele restricts his discussion of internalism to moral *cognitivism* (Mele, 1996).

'inverted commas' sense of belief (Hare, 1952:124-6). The temp agency manager doesn't believe that sex discrimination is wrong, but 'believes that sex discrimination is wrong,' where the latter phrase means 'believes that discrimination based on sex is in accordance with what other people take to be wrong.'

The difference between the person who believes in an inverted commas sense and one who really believes is, first, that of content: the two don't believe the same things. The person who believes in an inverted commas sense only knows what other people take to be pretty important, and recites that belief as his own. One can believe in an inverted commas sense if they believe insincerely, like the sexist temp-agency manager. Another results from lacking the phenomenological experience to grasp the relevant concept. Think of making claims about the color of certain objects under the condition of colorblindness. It seems that one who's never had color experience doesn't know what the difference is between red and orange. There seems to be a certain quality to color experience, one essential to the understanding of what color is. Yet the colorblind person can know what other people call 'red' and 'orange.' So if he calls an apple 'red', it seems that what he means is 'what is in accordance with what other people call red.'[26] He believes in an inverted commas sense of 'red,' but doesn't believe insincerely.

Do the disaffected sloth and the self-aware anorexic believe in an inverted commas sense? There seems to be little reason to think that they believe insincerely. The sexist temp-agency manager has some motivation to put on an act, but neither of these two unfortunate people do. Rather, they seem genuinely concerned with what their beliefs are about. We can even imagine each of these people coming down on themselves for how they feel. Imagine the

26 This is controversial. Brink (1997), for instance, disagrees with the assessment I present here. Others, however, take the present view. See, in particular, Smith (1994:68-70).

self-aware anorexic, for example, saying something to herself like, “I wish I didn't feel so ashamed of myself. I really have nothing to feel ashamed of!” Likewise, it's no stretch to suppose the disaffected sloth incriminating himself for being so inert as to not feel ashamed for what he himself judges shameful. But on the other hand, neither of these people are like the congenitally colorblind. They both know what feeling shame is like, and so would seem to have the experiential resources available to judge correctly that their situations are shameful.

These cases are anomalous in the sense that they involve people whose emotional state and corresponding belief state are systematically at odds due to what appears to be a psychological and/or physiological handicap. Both may be depressed, for example, and suffer from chemical imbalances that prevent their sentiments from corresponding appropriately to their evaluations and vice versa.[27] We might say that 'normal' people, i.e., anyone not psychologically or physiologically handicapped, would not be able to ignore either the motivating force of their value judgments or the constraints that their sentiments place on their evaluations.

These remarks suggest a further restriction to an analysis of value statements that has an analogy to color terms. Color vision, of course, has its anomalous cases as well, where physiological impediment (colorblindness) and circumstantial obstruction (poor lighting) may impede one's ability to refer correctly to color properties. Color properties can be given a dispositional analysis in this sense: an object is some color if and only if it appears to be that color to normal observers under normal circumstances. This move is legitimated by the fact that objects typically appear similar to observers who have relevantly similar physical conditions and

27 The term “appropriately” is vague here. D'Arms and Jacobson attempt to identify the precise meaning of the “appropriateness” of an emotion. I'll be turning to their discussion shortly.

who are in relevantly similar situations. Thus, those who don't see colors the way others see colors are those with jaundice, for example, or who are colorblind, or who make their observations in poor lighting, etc. All of these exceptions are relevantly similar conditions in that they obstruct what would otherwise be the way objects appear. An analysis of value statements might follow the same strategy, and offer the hypothesis that value properties are those that normal persons under normal circumstances would pick out. Those who suffer from weakness of will or some other psychological malady, as well as those who have not had the benefit of a sound moral education, would be eschewed from the analysis.

There are several problems with the analogy between values and secondary properties, however. First of all, those trying to discern the color of an object in poor lighting are probably going to fail. But we may wonder whether a perverse moral education will consistently produce the same aberrant results. Consider this unfortunate example. Suppose that someone, call him Jim, was brought up to think that cats were evil and that they should be destroyed. Jim has been around many cats and has seen his family do awful things to them. Yet, Jim also has a fondness for cats and feels guilty about what he and his family have done to them. Though it is not implausible to suppose that his upbringing explains the fact that his beliefs run contrary to his feelings, we could just as well imagine that his personal regard for cats would put pressure on him to change his beliefs. We could imagine Jim very likely not believing that cats were evil despite his moral education, and this suggests an asymmetry between his circumstances and that of the colorblind person trying to discern the color of an object. The dispositionalist must claim that circumstances akin to one's upbringing constrain one's value judgments or sentiments as

uniformly as poor lighting constraints on one's ability to differentiate color. And this seems implausible.

Of course, objects in dim lighting can be viewed normally after a period of adjustment. Say that Jim's moral education does not influence him to the degree that he believes that cats really are evil and that they should be destroyed. In this case, Jim's moral education may be like looking at a red object in dim lighting and eventually seeing that object as red; he's been taught to be disposed to treat cats with hostility, but is not so disposed. His natural response to cats can take form despite his moral education, much like one's view of red objects can be sufficiently discriminating to identify objects that are red in poor lighting.

But there are other disanalogies of moral properties to secondary properties like color. First, it seems that secondary properties and moral properties supervene on the world in much different ways (Blackburn 1985). The fact that color properties, for example, supervene on some base subset of physical properties is a scientific one. Someone could fail to understand this fact without being incompetent in their use of color concepts. I can know what red is without knowing or believing that red could not fail to exist in a world physically identical to ours. But the same is not true for the concept WRONG. If I'm a realist about morality, and if I think that killing an innocent person is wrong, I must think that it is wrong in a world that is physically identical to ours. The physical subset that constitutes the supervenient basis for having the property wrong is present in both worlds. So if it's wrong in one world its wrong in another. Unlike color concepts, I cannot fail to understand this fact about morality and competently use moral concepts.

A related concern can be brought out by emphasizing the sometimes dissonant relationship between what we think is wrong to do and what actually is wrong to do. On a dispositionalist account of color, the world would cease to be red if it ceased to appear red. And this can happen if we change sufficiently. But in the moral case properties do not change with those who observe them. A dramatic change in the way we respond to the killing of innocent people would not change the fact that killing innocent people is wrong to do (if there is a fact of the matter). Thus, we are inclined to say that even if everyone condoned the killing of innocents, it would still be wrong to do. But this is not true of the color case. If every one changed so that no one any longer saw red, then there would be nothing that was red.

Another disanalogy is the distinctive tendency of moral dispute to become intractable. Two people can agree to all of the natural facts of the matter and yet disagree as to what or whether there is anything of any value. Return to the dispute between deontologist and consequentialist over whether the killing of an innocent person is wrong to do. Suppose that by killing one innocent person the lives of thousands will be saved. The deontologist says that it is wrong since killing an innocent person violates standards of blame. The innocent person is blameless, and so his right to life cannot be revoked. The consequentialist says that it is right because of the thousands of lives that will be saved. Both are aware of how many lives are at stake, and of our beliefs about his innocence, yet neither will be pulled away from their normative convictions. They disagree because they have different intuitions about what it means to do something wrong. The deontologist believes that wrongdoing consists in violating a moral rule, while the consequentialist believes that wrongdoing consists in producing a result that contributes to an optimal measure of pleasure, or preference satisfaction, or whatever.

Employing different concepts prevents each from being able to convince the other that their take on the issue is correct, since neither can give a reason for thinking that killing an innocent person is wrong that the other will be satisfied with.

Disputes over color are not like this. Cultures that emphasize certain colors can notice them with more acuity. If different shades of red permeate the environment it may cause the people who live in it to notice different shades of red with regular success. However, different cultures who do not share this distinctive color vision will not challenge the others' use of the word 'red.' Moral concepts, on the other hand, appear to be essentially contestable. It seems that it is in principle possible for two to share in all of the relevant facts of the matter and disagree about whether Φ-ing is wrong.

Of course, there may be ways around these objections. Differences between value and color may only motivate the need to produce a dispositional account of value that is sufficiently different from color. And there are likely other ways to develop a dispositional account of value that has little in common with the kind of perceptual model we have to deal with color concepts. One might speak of being disposed not to see wrongness *per se*, but just to have some *pro* or *con* attitude. Or perhaps a sentimentalist account is forthcoming. Perhaps something can be made of the idea that value judgment is just being disposed to some kind of emotional response. These are legitimate alternatives.

Nonetheless, the dispositional analysis has difficulty capturing the normative force of value terms. Concepts like SHAMEFUL and DISGUSTING play a role in governing the sentiments that we have. Yet the dispositionalist has nothing to say about why value statements in standard contexts are the ones we should defer to. Proponents of this insist too strongly that outlier cases

are either incoherent or insincere, while disagreement with the norm is all too common and compelling. Consider an example given by D'Arms and Jacobson:

> Consider the heretical view that the quintessential American delicacy, the Big Mac, is in fact disgusting. A dispositionalist might try pointing out just how many billions of them have been sold worldwide, but that would be to no avail. The heretic does not doubt that most people love them; it's just that this fact fails to move her. Evidently, most people's taste is abominable. Just look at the thing, she might add, all fatty and processed in its cardboard box, dripping with 'special sauce.' If you don't see what is disgusting about that, so much the worse for you. (2000a:727)

Nothing appears to be wrong with the heretic's view. The average American's regard for the Big Mac is wrong, she says, despite the fact that he may be quite normal physiologically and psychologically, and have had the experience of a perfectly normal upbringing. The dispositionalist must meet this objection on normative grounds. An open-question argument looms here: the fact that normal people under normal circumstances make certain value judgments doesn't mean that those judgments are right.

The dispositionalist might respond by arguing that the vast majority of Americans are not actually normal people. Though there doesn't seem to be much evidence that we are psychologically or physiologically impaired, perhaps something can be made of the suggestion that our views and tastes have been corrupted by the fast food industry. The influencing affects of advertising, the convenience of the drive-through, the accommodations made for children—

such as the play scape and the use of clowns and toys—and even the relative inexpense of the product have likely played a role in shaping Americans' attitudes toward the Big Mac. Perhaps under circumstances absent these influential ploys, the vast majority of Americans would in fact find the Big Mac disgusting.

The problem with this is that it just doesn't seem true. We can suppose that even if most people were accustomed to a diet that generally excluded fast food items, many would still not find the Big Mac disgusting. Fatty and salty foods, after all, taste good. It is perhaps the biggest selling point of the fast food industry that their products are so tasty. They are tasty not only to those who have an unusual gustatory apparatus but to 'normal' people. This scenario is not unlike that of Jim's, in which a substantial possibility of his believing contrary to the views proposed by his moral educators can all but be ruled out. Indeed, we may be naturally disposed to enjoy the Big Mac, and need little help from Mc Donalds to form a favorable assessment of it.

The difficulty here is one of reduction. The fact that a certain class of people under certain conditions approve of something doesn't, by itself, appear to have any normative implications. More generally, things we value, things we call 'good' and 'bad,' 'right' and 'wrong,' are things we take to merit the status. Normative assessments are not simply descriptions of the natural world but recommend something of practical import. It seems that we can always ask whether a natural object is 'good,' for an assessment of its natural properties leave this question open (Moore 1903).

I don't mean to suggest that it's impossible to bridge the normative gap between value and nonvalue terms. But it is no small advantage to any view that can avoid this reductive task.

A rational version of sentimentalism does just this. It avoids reduction by rephrasing value claims in terms of deontic claims about what reasons there are to have some response.

1.3 The Response Dependency Thesis and 'Fittingness'

RDT is intended to capture the special normative aspect of our value judgments. It says that evaluations are equivalent to judgments that a certain emotional response is fitting, or appropriate.[28] However, these terms (fitting, appropriate) are elusive.

Even the term 'evaluation' is ambiguous (Brandt, 1946:110).[29] Consider first adjectives that denote the property of having a certain kind of effect, like 'beneficial' and 'edifying.' These terms describe an object in such a way as pertains to one's well-being. Another group is comprised by terms which express conformity to or deviation from some standard, e.g., 'decent,' 'perfect,' 'lacking,' and 'correct.' Another group of evaluative terms refers to some emotion or attitude, as in 'disgusting' and 'shameful.' RDT is concerned only with this latter class of evaluative adjectives.[30] We don't often mean anything of moral significance when we say something like "X is perfect," but we often do when we say "X is shameful."

---

28 D'Arms and Jacobson prefer the former term. For the sake of convenience, I will use both terms interchangably.

29 Brandt offers what must be close to a complete taxonomy of evaluative terms (1946:109-113). He also distinguishes evaluative adjectives from, what he calls, "trait names," which specifically characterize one's trait, as in "abject, abrupt, absent-minded," etc. Evaluative adjectives may imply that one has a certain trait, but they do not specifically denote traits (1946:107-109).

30 Brandt is credited with being the first to explicitly pick up on this idea, which neosentimentalists incorporate and develop later on. See D'Arms and Jacobson (2000b:6): "[Brandt's] suggestion that our emotional responses can be assessed for their fittingness is the cornerstone of a metaethical program that has reinvigorated the sentimentalist tradition in ethics..."

For simplicity's sake, I'm going to restrict the discussion here to the emotions, though other affective states might fall under the (RDT)'s scope and require separate treatment. For one author's take on the differences between various affective mental states, see Prinz (2004a:179-197).

What is common to all of the terms that refer to an emotion is that they seem to convey a meaning synonymous with 'being worthy of,' or 'deserving of.' The idea is that, as Brandt puts it, "the attitude or emotion [is] properly, rightly, or fittingly directed at the object (1946:111)." Other evaluative terms do not have this denotation. We don't mean, for example, that by 'X is correct' that 'X is worthy of correctness,' or that by 'X is beneficial,' that 'X is deserving of benefit.' Typically, the evaluation that bears some moral significance have a variety of related suffixes, such as *-ing, -ous, -ful,* and *-ive*. But this isn't a rule that allows us separate morally neutral evaluative utterances from those that are morally charged. Nonmoral adjectives sometimes share in these same qualities. Consider the term 'shocking,' which expresses the idea that an object is 'deserving of shock' (not in the object but in the observer), and has an *-ing* suffix. Still, these rules can be helpful in parsing out the relevant evaluative terms.

In a similar fashion, we can get some hold on the meaning of 'fitting' or 'appropriate' by process of elimination. Consider the following. Mary has trained for a particular position in an accounting firm for several years. She and her peers consider her most qualified for this position and so she gets the job. But after several disappointing progress reports she is laid off from work after only a short time. Due to her substantial training, Mary believes that her dismissal is shameful. The impact of this belief is considerable, and Mary soon sinks into a deep depression. Unfortunately, Mary also has two small children who are relying on her to provide for them. But Mary simply has no motivation to go out and find new work. Mary is aware of this, and believes that if she were not ashamed of her failure she would not be so depressed. Further, she believes that by eschewing her feeling of shame would she significantly increase her chances of finding another job. Yet, the shameful feeling of being previously dismissed from work prevents her

from taking any action. Mary therefore believes that her shame is fitting, but also that her shame is morally wrong to feel.[31]

The appropriateness of an emotion might also be conflated with its practical upshot. Consider an example taken from D'Arms and Jacobson:

> Dennis is a graduate student in philosophy who is ashamed of his inability to articulate his views clearly—a sentiment we can all too easily recall. Moreover, he thinks this shame appropriate, in the relevant sense: it's shameful to be such an inarticulate philosopher. Yet he realizes that philosophical discourse is a skill one improves with practice and that his shame inhibits him from speaking up in seminar, thus exacerbating the problem. Best not to be ashamed, then, if he can help it. Insofar as his emotions are under voluntary control (which, if he is like most of us, is not very far), he will not be ashamed of his Latin malapropisms and clumsy attempts at counterexample. So Dennis seems not to endorse feeling ashamed of his abilities; indeed, he apparently endorses not feeling shame. But he does not thereby deny that his inabilities are shameful. (2000a:741)

31 I assume that an emotion may be morally wrong to feel if it would lead to behavior that would violate some moral obligation. This claim is controversial, of course, for it suggests that feelings are like actions, where assessments of moral rightness or wrongness are connected to judgments of responsibility—and it is questionable to what degree we are responsible for the feelings we have. Nevertheless, we do often evaluate a person's character on the basis of her emotional responses; and we typically hold people responsible for their characters when they have had a sufficient amount of time to establish a virtuous disposition. Moreover, it seems that the moral status of an action need not depend on whether one is responsible for that action. That is to say, it seems that one could do something wrong without being responsible for doing it. I think that the same case could be made for the emotions.

These examples are not uncontroversial. There may be good reason to think that the moral inappropriateness of an emotion does bear on its fittingness. But this is not the place to settle such disputes once and for all.[32] It seems reasonable to make these distinctions. There appear to be plausible cases which exemplify the difference between an emotion's fittingness and it's practical or moral implications.

As the examples of Mary and Dennis suggest, fit appears to have nothing to do with how an emotion will affect the agent, which is precisely what determines an emotion's moral or practical quality. Instead, fit has something to do with the relationship of an emotion to its object. Both Mary and Dennis judge their shame to be fitting because they have certain beliefs about their own inadequacies or failures. The implication is that the inability or failure to live up to what one considers a reasonable ideal merits feeling shame. By contrast, imagine that, instead of shame, Mary had felt joy for having been laid off in her particular circumstances. I think we could say that her joy is inappropriate. And suppose that, instead of shame, Dennis had felt pride about his inability to express his views clearly. Again, his pride seems unfitting. What seems to matter to the assessment of their emotional state, in regards to its fit, is not whether that emotion would be bad or unhelpful to feel, but whether they or their circumstances really merit feeling it.

Parallel support for this idea can be found in literature in the philosophy of emotion, which suggests a well accepted view according to which emotions are mental states that represent objects as having certain value-laden features.[33] An emotionally significant object

32 For a thorough discussion whether moral considerations bear on an emotion's fit, see D'Arms and Jacobson (2000b).

33 It is misleading to describe the situation in the philosophy of emotion as one of convergence and agreement, however. Though many hold the very general view stated above, there is notable disagreement over the cognitive status of emotions, as well as to how we should analyze them. Solomon (1976), for instance, believes that emotions are judgments of a special sort, while theorists such as Roberts (1988) and Prinz (2004a) disagree. Yet Roberts and Prinz differ on what methodological treatment to use. While Roberts operates via conceptual analysis to determine the proper objects of emotions, Prinz understands their representational content in terms of

could be anything. It could be a situation, such as being late for a meeting. It could be a thing, person, or animal. It could be an event, such as being fired or ridiculed. It could even be an action, as when one lies or steals. So the emotion of fear, for example, is said to represent those features of an object in virtue of which it is dangerous, sadness to indicate the occurrence of a loss, anger the event of some offense, etc. Philosophers have used these various object kinds to individuate different emotions, as well as to accommodate the idea that emotions have the property of intentionality, and that they are in some sense meaningful or especially important to the agent (Solomon 1976, Roberts 1988, Lazarus 1991, Prinz 2004). Anger, for instance, involves being angry at something or somebody; it has an intentional object. But this object is tied to a notion of having committed some offense, which is necessary to construe the sense in which the emotional object is meaningful to the one who's angry.[34]

Different emotions have different kinds of objects to which they are fit. Suppose that instead of shame, Mary feels joy at having been fired, or that Dennis feels proud about his inability to express his views clearly. Both Mary and Dennis's emotions seem inappropriate because they don't appear to correspond to the right kinds of objects. There's some play in determining which emotion is fit to which kind of object. If Mary were to feel sad about losing her job, for example, this wouldn't seem so odd. One is normally sad after incurring a loss. By the same token, some emotions seem to have a rather wide application. Shame, for instance, is normally self-directed, i.e., we normally feel ashamed about ourselves. But we also often speak of being ashamed of someone else, as when a mother says to her child that she is ashamed of him. So the idea that emotions have kinds of objects to which they are fit should be tempered by

---

the ecological role they play in our natural and social lives. A notable exception to the above claim is Griffiths (1997), who argues that emotions are not amenable to any unifying treatment.

34 Thus, Roberts (1988) describes emotions as 'concerned based.' Prinz (2004a) too, uses similar terminology.

the recognition that emotion categories are vague. But this does little to debunk our right to criticize an emotion's fittingness if it doesn't correspond, roughly, to the right kind of object.

Against this theoretic backdrop there emerge two individually necessary and jointly sufficient conditions for fittingness, one of which has already been explained. First, an emotion must represent the relevant value-laden features of its object.[35] Consider again Mary, who feels joy because she's just been fired from a job she was more than qualified to do. Mary's joy is unfitting because there is nothing about her circumstances to feel joyful about. Her failure to secure employment simply lacks the relevant evaluative features that the object of joy should have. She has not enjoyed a pleasant surprise, or benefited from an unexpected visit from a friend. Rather, she has suffered something that should be a terrible blow to her self-confidence.

Second, an emotion must represent an object that actually has the relevant value-laden properties. For example, if someone, call her Sally, is afraid of what she believes to be a coiled black snake in her closet, but which is really a pile of black socks, her fear would be unfitting. This kind of misrepresentation would also include instances in which the emotionally significant object just doesn't exist, as in being envious of something someone really doesn't have. In each case, there is not just a failure to correctly portray an object's value-laden properties, but a failure to correctly refer to an object at all.

This second condition may not seem necessary. One might argue, for instance, that Sally's fear is actually appropriate, since she is warranted in thinking that there is a snake in her closet. Let me flesh this example out a bit to make it a little clearer. Imagine that Sally didn't know she was mistaken. Suppose also that she believes correctly that she has nearly perfect

35 D'Arms and Jacobson call this a requirement of an emotion's "shape." As they put it, "an emotional episode presents its object as having certain evaluative features; it is unfitting on grounds of shape when its object lacks those features (2000b)." The second condition too is a requirement of shape.

vision, and that there's decent lighting in her closet. In these circumstances, Sally's fear is surely warranted; there's no consideration that could come to bear in undermining what she sees. Compare this case with someone else, call him Larry, who is terribly afraid of flying. To alleviate his fear, Larry has learned a great deal about airplanes and has collected some reliable statistics on the safety of flying. In light of his research, Larry believes that flying is a safe means of transportation. But whenever he steps onto a plane he becomes mortally afraid of the thought of crashing to the ground. Larry's fear, unlike Sally's fear, seems inappropriate. Not only does he believe that he has no reason to be afraid, his belief is based on reliable and conclusive evidence. This suggests that the appropriateness of an emotion depends on what one is subjectively justified in feeling, given one has sufficient evidence to believe that some object has the relevant value-laden features.

The problem with this suggestion is that it's false. RDT states that to think that X is Φ (some evaluative property) is to think that F is an appropriate emotional response to X. On the current suggestion, RDT says that thinking that X is Φ is equivalent to thinking that F is a subjectively justified response to X. But put this way, RDT is false. Imagine that there really is a black snake in Sally's closet, but it's hidden under a pile of black socks. Sally has no reason to believe that her situation is dangerous, and is subjectively justified in not being afraid of anything. But this doesn't mean that she isn't in a dangerous situation. The question of whether an object is Φ and whether one is subjectively justified in feeling F in response to that object diverges in cases where evidence is lacking or misleading. So the question of whether X is Φ is not going to settle the question of whether F is a subjectively justified response to X. Fittingness, rather, is constrained by the evaluative properties of X.

The oddity of feeling an emotion for reasons that speak to its prudential or moral upshot suggests something special about fittingness reasons. They are, we might say, the right kinds of reasons for feeling an emotion. One who does not feel what's fitting is subject to a kind of criticism that is not available when appealing to a moral or prudential basis. Notice how there is something particularly convincing in saying that Mary should be ashamed for failing at her job. It may not be in her best interest to be ashamed, nor what she should feel in a moral sense. But we would be likely to forgive Mary if she felt ashamed even in light of having more reason not to feel ashamed. She might be irrational in the sense of not feeling what she has most reason to feel, but not abnormal. Her response is what we would expect of anyone disposed to feel ashamed of doing what they think is shameful. In this way, fittingness reasons make sense of our responses more than moral or prudential ones. They appear to speak to an emotion's logic, if you will.

The peculiar normative force behind fittingness reasons raises the question of how we might distinguish reasons of the right kind from those of the wrong kind. Making sense of this distinction is important, since fittingness reasons do appear to warrant some privilege among others, while on a par with other kinds of reasons about what to feel, all things considered. There should be some way to make sense of this distinction while accommodating the fact that the wrong kinds of reasons can perfectly well justify feeling F.

1.4 The Wrong Kinds of Reasons

Moral and prudential reasons justify feeling an emotion, but they're the wrong kinds of reasons. Emotions can fit their objects while not being morally right to feel or being in one's best interest to feel. Reasons that bear on an emotion's fit are the right kinds of reasons. But what is it about these reasons that makes them the right kind? What makes a reason the right kind of reason?

The difficulty involved in explaining the distinction between the right and wrong kinds of reasons has earned the task labels such as the 'conflation problem' or the 'wrong kinds of reasons problem' (D'Arms and Jacobson 2000, Rabinowicz and Ronnow-Rasmussen 2004, Hieronymi 2005). According to the standard view about what reasons are, we seem unable to distinguish between different sets of considerations that count in favor of certain attitudes, only one of which is the right kind. The standard view has it that a reason is a consideration that counts in favor of an action or attitude (Hieronymi 2005). As such, moral, prudential and fittingness reasons are indistinguishable in their ability to justify an attitude.[36]

36 There's an important distinction to make between two different kinds of reasons, ones that explain and ones that justify. Suppose that Sally believes that there's a snake in her closet and doesn't want to open the closet door. Sally's belief that there's a snake in her closet explains why she's afraid when she opens the closet door. Suppose further that there really is a black snake in her closet. The content of her belief 'there's a black snake in the closet' would justify her fear. Accordingly, we can read reasons as mental states, such as beliefs, desires, emotions and so forth, that might explain an action of an agent, or as the contents of these mental states. The contents are commonly called 'considerations.' Since fittingness is a deontic notion, the account is formulated in terms of considerations that bear on an action or attitude to indicate that the reasons in question here are justifying ones, not explanatory ones.

There is also an important difference between having a good reason for having some attitude and taking oneself to have a good reason for some attitude. Sally believes that there's a coiled black snake in her closet. But now suppose there really isn't. She may be justified in believing there is, but in fact what she takes to be a black snake is really just a pile of black socks. Sally thus takes herself to have a good reason to be afraid even though in fact she does not. Sally's reason for her fear is *prima facie* justified. But her reason to be afraid is not a good one. In some parts of this discussion I'll refer to good reasons. These kinds of reasons bear on an emotion's fit, since an emotion is fitting only if it correctly represents its object. In other parts of the discussion I'll refer to *prima facie* or operative reasons. These are reasons that would justify having some attitude in a subjective sense, but they are not ones that bear on an emotion's fit, since one could believe oneself to have good reason to feel some emtion and be wrong. An emotion is not fitting in these kinds of cases. In order for an emotion to fit its object, that object must have the relevant value-laden properties.

The problem, then, is that if reasons are considerations that count in favor of an action or attitude, how can we distinguish those wrong kinds of reasons from the right kinds? Both moral and prudential reasons, as well as ones that bear on an emotion's fit, are reasons that count in favor of having an emotion. But only the latter are of the right kind. An account which holds that reasons are simply considerations that count in favor of an attitude or action is therefore unable to provide the resources necessary to distinguish between reasons of the right kind and reasons of the wrong kind.

1.4.1 Object and State-Given Reasons

One way of tackling the wrong kinds of reasons problem is to suggest that the wrong kinds of reasons really aren't reasons to feel an emotion at all. Consider Parfit's distinction between 'state-given' reasons and 'object-given' reasons. As he puts it,

> Of our reasons to have some desire, some are provided by facts about this desire's object. These reasons we can call *object-given*. We can have such reasons to want some thing either for its own sake, or for the sake of its effects.... Other reasons to want some thing are provided by facts, not about what we want, but about having this desire. These reasons we can call *state-given.* (2001:21-22)

According to Parfit, a reason to have some desire that is based on whether it would be good to have (or morally right to have) are really reasons just to want to have it. State-given reasons,

then, are really object-given reasons to want to have a desire. When we believe a desire to be in our best interest we take that property of the desire to justify our wanting it. These are not reasons to have the desire, but rather to want to have it. Gibbard makes a similar suggestion by pressing the difference in what he takes to be the rationality of a belief and its desirability. He gives the following example: "it might be disadvantageous for one of Cleopatra's courtiers to be angry at her, even if she ordered an execution unjustly, and it thus 'made sense' to be angry at her. For the courtier might want to ingratiate himself with her, and he might rightly fear that anger would cloud his countenance and spoil his charm." In such a case, he says, "it is rational *to be angry*, but also rational *to want not to be angry*" (1990:37, original italics).

The reasons that speak to the desirability of an attitude, according to this view, are not reasons that justify those attitudes. Thus, the distinction between the rationality of an attitude and its desirability makes the wrong-kinds of reasons problem disappear. It neutralizes any distinction to be made between appropriate emotions where some are of the right kind and others are not. There are no 'wrong kinds' of reasons, on this view, because such reasons do not justify the emotion at all. They are reasons not to feel it, but only to want to feel it.

This suggestion seems plausible, but on a closer look it's unconvincing. Consider Mary again. Mary feels ashamed for having been fired from her job. She thinks her shame is justified, in the sense that her circumstances call for feeling shame, but also believes that her shame is counterproductive. According to the above view, Mary has no reason not to feel shame. She has a reason to want not to feel shame, but no reason not to feel it. And this is implausible. The fact that Mary's shame would be debilitating seems like a perfectly good reason not to feel shame. Consider how we often justify our actions by appealing to their moral or prudential implications.

Proponents of the above view might insist that these reasons are not actually reasons that justify doing that action either, but only justify wanting to do that action. Reasons that justify an action would seem, in this case, to derive from the circumstances in which the action would take place. So the action could be called for by the circumstances and justified in this way, while its moral and prudential implications bear only on whether to want to do it. Surely, however, the moral and/or prudential upshot of doing some action is a reason to do or not to do it. Why don't these reasons play the same role with respect to our feelings? Why discount those kinds of reasons here? To these questions Gibbard and Parfit appear to have nothing to say, and it seems they just have different intuitions about the matter.[37]

Still, perhaps something can be made of the distinction between object and state-given reasons. We might say that the right kinds of reasons derive from facts about the objects of one's emotion, while the wrong kinds of reasons derive from facts about the emotion itself. This too seems natural, I think, because moral and prudential reasons derive from facts about the emotion itself. The fact that an emotion would be morally wrong to feel does not show that it is fitting. Neither does the fact that an emotion would be prudential to feel. Perhaps this is because they derive from qualities of the emotion rather than the object of the emotion. The suggestion here then is that the right kinds of reasons derive from objects of the emotion, while the wrong kinds of reasons derive from the emotions themselves. This would make the object/attitude distinction exclusive and give us a way to distinguish the right kinds of reasons from the wrong kinds.

A problem with this suggestion, however, is that object-given properties can provide reasons to have an attitude that are of the wrong kind. Consider first the admittedly strange

37 Others make a similar complaint. See D'Arms and Jacobson (1994), and Rabinowicz and Ronnow-Rasmussen (2004).

attitude-mirrored properties of objects: for any property of an attitude P, there is a corresponding property of an object, P', that any attitude towards it would have property P. Attitude-mirrored properties of objects, it would seem, provide reasons of the wrong kind. In the case of Mary's shame, for instance, we can say that her feeling is crippling or disadvantageous in some way, and that the object of her shame—having failed at her job—is such that her shame about failing at her job is crippling. Thus, the object of Mary's shame has a property that could be taken as a consideration that does not bear in the right way towards whether Mary's shame is fitting.[38]

This might compel one to restrict value to the intrinsic properties of an object. Unlike the relational property of an object being such that a certain attitude towards it has some property, intrinsic properties are context independent. We might understand an object's value in terms of some set of intrinsic properties that it has, or perhaps as itself an intrinsic property of an object (Moore 1903). In this case, any reason deriving from an intrinsic property of an object would be of the right kind.

But this view also has a few problems. First, it requires a response-dependent analysis to focus solely on final value, when some things appear to have value only instrumentally. Imagine someone who works for a corrupt firm. Say the work he does is not itself wrong or underhanded in any way, but it contributes to the firm's ability to do things that are wrong and underhanded. This person should feel guilty about what he does, not because his work is itself wrong but because he contributes to activities that are wrong. His work is not intrinsically but instrumentally wrong. Second, even if there's a way around this obstacle, it seems contentious that all final value must be intrinsic. A work of art, for instance, may have final value at least in

38 This problem is most likely no big deal in any event. At the very least, there seems to be consensus that attitude-mirrored properties are controversial. It's questionable, for instance, whether we should worry too much about properties that don't have any causal powers. See Shoemaker (1980).

part because it is unique or because it was made by a certain artist. Likewise, facsimiles do not have the same value as the genuine articles. The value they have is derivative of how they came to be, not something intrinsic to them. Relational properties therefore do seem capable of comprising the final value of an object, and probably shouldn't be ruled out by an account of value.

Here's another way how the intrinsic properties of an object could provide us with the wrong-kinds of reasons for valuing it. Imagine an evil demon who demands that we admire him because of his disposition to punish us if we don't (Rabinowicz and Ronnow-Rasmussen, 2004:419). The demon's disposition to punish us is an intrinsic property that he has. Admiring him on this score would be to admire him for his own sake. But the demon is in no way admirable for being so disposed. Admittedly, there's a question whether admiration is psychologically possible in this instance. It does seem odd to admire the demon for the reason the authors give. Fear seems more fitting. But it doesn't take that much to imagine that one admires the demon because he's so malicious. This kind of perverse attraction to evil might even be augmented by the fact that the demon is disposed to punish us, rather than someone else. Masochistic admiration of the sort likely exists. In any event, the attitude isn't incoherent, and should be admitted as a counterexample to the claim that the intrinsic properties of an object will provide only reasons of the right kind to have some attitude.

Olson has recently set out to fix problems like these by offering an account that distinguishes reasons of the right kind from those of the wrong kind based on the object/attitude-given distinction. He offers the following schema:

> "an object *O* is valuable iff you ought to ([A], [O], [JE]), where the *A*-clause is a placeholder for the relevant attitude, the *O*-clause specifies the object in question, and the JE-clause gives the justificatory explanation of why you ought to take up *A* towards *O*. (2004:298)

As an ought statement, the schema would be paraphrased: you ought to [A] towards [O] because [JE]. Olson states that JE-clauses must not be A-referential: they cannot refer or be about the attitude in question at all. As he puts it, "JE-clauses must not contain any reference whatsoever to properties of the attitude in question, whether in the guise of properties of the attitude or, more limitedly, of the object" (2004:299). This restriction, he supposes, will filter out those reasons of the wrong kind.

Consider first an example of an ought-statement that prescribes having an attitude for the wrong kinds of reasons, the statement that Mary ought not to be ashamed of having lost her job because it would be emotionally crippling. This would be analyzed: Mary ought not to feel [shame] about [failing at her job] because [shame is crippling]. The JE-clause is A-referential. It refers to a property of her shame to justify that feeling. Accordingly, Mary's shame is not fitting because it is not supported by a reason of the right kind. Now suppose that Mary ought to feel shame because she failed at a job she was adequately trained to do. The ought statement would now read: Mary ought to feel [shame] about [failing at her job] because [she was adequately trained to do her job]. The JE-clause in this statement is not A-referential, and so is a reason of the right kind.

Olson's schema also rules out justifying feelings based on properties of objects that mirror properties of attitudes. Using the example of Mary, we get: Mary ought not to feel [shame] about [failing at her job] because [failing at her job is such that being ashamed of it would be crippling]. Here the JE-clause refers to properties of the attitude in question, and so is a reason of the wrong kind.

Finally, the schema works with the troubling example Rabinowicz and Ronnow-Rasmussen give. Translated into an ought-statement, we get: you ought to [admire] the [demon] because [he is disposed to punish you if you don't]. Again, the JE-clause makes reference to the attitude in question, at least implicitly. It says that the demon is disposed to punish you if you don't admire him. So it's a reason of the wrong kind.

But here's a problematic case. One ought to love romantically, in the sense of being appropriate or fitting, only those who respond well to one's love (Hieronymi 2005:447). Using proper names, say that Tom ought to love Mary because she responds well to his love. On Olson's schema, the ought statement would read: Tom ought to [love] [Mary] because [Mary responds well to his love]. Here, the JE-clause makes reference to Tom's attitude, but is a reason of the right kind. Thus, the right kinds of reasons can be A-referential.

This last example may seem controversial. The fact that Mary is responsive to Tom's advances is surely a reason for him to pursue her, but it may be of the practical sort. Tom may enumerate all of the wonderful features that Mary has but her responsiveness may not be among them. 'She's great,' he tells his friends, 'but she's not interested.' Mary is lovable in Tom's eyes, and the fact that she doesn't love him doesn't make her less lovable.

However, though one's responsiveness is a practical reason to love someone, it is also a reason that counts towards being lovable. It's not uncommon that we explain attraction by citing a favorable response. The better that Mary responds to Tom, the more attractive she will appear to him. Thus, it seems fair to say that she is a fitting object of his love in part because she responds well to it. By contrast, if Mary were completely unresponsive to Tom his love may be unfitting. Say, for example, that Tom is stalking Mary when she doesn't want to have anything to do with him. His love is unfitting here, despite the fact that Mary has many lovable characteristics. She desires to be away from him, and it seems odd to love those who want nothing to do with you.

These attempts at making the object/attitude distinction exclusive fail. Reasons that derive from the object of one's attitude can bear on whether it is good to have that attitude. And these can be the wrong kinds of reasons. Reasons that derive from one's attitude can also bear on whether the object of one's emotion is fitting, as in the case of Tom and Mary. In the next section, I examine another proposed solution to the wrong kinds of reasons problem that makes the object/attitude distinction exclusive.

1.4.2 Constitutive Reasons

The shortcomings of views that utilize the standard view of reasons may encourage a different approach. Hieronymi has recently done this. On her view, a reason is a consideration that bears on a question. This hypothesis is born from general observations about reasons and how they function in argument:

> A piece of reasoning has some direction to it—it is the sort of thing that might have a conclusion, in something like the way an argument has a conclusion. And the items in a piece of reasoning...stand in certain relations to each other—they bear what we might call "rational relations" to one another—and thereby also bear on the conclusion. Further, reasoning seems called for just in case a conclusion is unknown or called into question. So I suggest, for consideration, the following account of a reason: a reason is a consideration that bears on a question. (2005:444)[39]

There are two ways in which a consideration can bear on a question. A consideration might bear on whether *p*, where *p* is the content of a belief, or whether belief that *p* is good to have. Consider the difference in an evaluation of the evidence and arguments for god's existence, and the result of weighing the beneficial and detrimental effects of believing that god exists. The former consideration is 'content-related' in the sense that it bears in some way on the content of the belief that 'god exists.' The latter consideration is 'attitude-related' in the sense that it bears on whether it is good to believe that god exists. Each counts towards an answer to a separate question.

Like the ambiguity expressed in the object/attitude distinction, content-related reasons and attitude-related reasons can bear on each other's question, as it were. Consider the fact that true beliefs are often good to have. Here, the question of whether a belief is true could bear on whether a belief is good to have. So considerations that bear on whether a belief is true can also

39 Hieronymi makes her case primarily in terms of belief. She intends to solve the wrong kinds of reasons problem on a structural level. So the fact that she talks mainly about belief only suggests that it's a convenient example to use.

bear on whether it is good to have that belief. The reverse is also true. Sometimes beliefs that are good to have are more likely true if they are believed in. Believing, for example, that one will win a game could make it more likely that one will win. In this case, considerations that bear on whether one will win are those that bear on whether it is good to have the belief that one will win. There are also odd cases where the content of a belief directly bears on whether it is good to have, for instance: the belief that 'this belief is good to have.'

The general point applies also to intentions. Consider a case where someone is trying to decide whether to go to the store or to the library. Say she considers the matter carefully and decides to go to the store. Her reasons for going to the store also bear on the question of whether it is good to have the intention to go the store, since going to the store probably requires intending to go. Reasons that count in favor of Φ-ing may also count towards intending to Φ, since intending to Φ is normally a precondition to Φ-ing.[40] As well, reasons that bear on whether an intention to Φ is good to have can also bear on whether to Φ. Think again of this person who's trying to decide whether to go the store or to the library. Imagine that she's really busy with her work while trying to make up her mind about whether to do one or the other of these things later in the day. It would be good, then, if she could go ahead and decide now that she will later do one thing rather than another. So her decision about what to do depends here upon the consideration that counts in favor of intending to do it.

We need some way of making the content/attitude-related distinction exclusive. Simply putting the matter in terms of considerations that bear on whether *p*, and whether it is good to believe that *p*, for instance, does not do the trick. Considerations that bear on whether *p* can also

40 A problematic case would be one in which there is reason to do something only if you do it unintentionally (Hieronymi 2005:446).

bear on whether it is good to believe that *p*. We need a way of construing reasons that allows us to specify those reasons that bear only on the question of whether *p* (*mutatis mutandis* for whether to Φ).

Hieronymi takes up the task by appealing to what she calls 'constitutive' reasons. These indicate a precise way in which we come to an answer to the question of whether *p*, something appropriately called 'settling' a question. Here's an example. Suppose that I'm trying to figure out if the mail came today. I notice that it's four o'clock in the afternoon, the time the mail usually comes. But I'm still not sure, and so go out to check the mailbox, and in it I find mail that was not there the day before. This piece of evidence 'settles' the question of whether the mail came today. It settles the question of whether *p*, an event that constitutes forming the belief that *p*. Constitutive reasons do not themselves constitute the belief that *p*, but rather settle the question of whether *p*, the settling of which constitutes the belief that *p*.[41]

All remaining reasons do not settle this question. They are oblique to the question of whether *p*, and bear on a different question: whether it is good to believe that *p*. For example, suppose that someone offers me $1,000 to believe that the mail came today. This is a good reason to believe *p*, but it is not a reason that would settle the question of whether *p*, and so would not play a role in constituting that belief. Rather, it is a consideration that bears on whether it is good to believe that *p*. Suppose I really need $1,000. The offer I receive could settle this question, which would constitute this second-order belief: the belief that the belief that *p* is good to have. Extrinsic reasons do not settle the question of whether *p*. They settle the

41 Cases of so-called 'akratic belief' are not counterexamples to this claim. As I understand it, belief is akratic when one bears in mind considerations that one takes not to settle the question of whether *p*, but belives that *p* anyway. Suppose I think that Jones' appearance indicates that he's trustworthy, even though I know that this doesn't mean that he is. I may still believe that he is when I'm around him, since his appearance is so convincing. See Scanlon (1998:35-36).

question of whether the belief that *p* is good to have. They too are constitutive, but they are not constitutive of the right attitude.

The right kinds of reasons for belief, then, are reasons that bear on the question whether *p*. Since reasons that bear on whether the belief that *p* would be good to have can also bear on whether *p*, they can be counted among those of the right kind.[42] Extrinsic reasons, by contrast, are all the other reasons that settle only the question of whether it is good to believe that *p*. These are the wrong kinds of reasons.

Attitudes supported by the right kinds of reasons are 'commitment-constituted,' as Hieronymi puts it:

> whenever one has an attitude that can be formed or revised simply by settling for oneself a question or set of questions…one is committed to an answer to the relevant question(s). One is committed in the sense that, if one has the attitude, one is answerable to certain questions and criticisms—namely, those questions or criticisms that would be answered by the considerations that bear on the relevant questions. (2005:450)

42 One might think that the consideration that the belief that *p* would be good to have is the wrong kind of reason, despite the fact that this very belief makes it more likely that *p*. Considerations that bear on whether one has reason to believe that one will win a game, for instance, are things like having a strategy that works most of the time in relevantly similar circumstances, the fact that one is cleverer than one's opponent, and a host of other things. But the fact that one believes that one will win is just a mental trick. It's not a real reason to think that one will win, just a way of psyching oneself up to win.

But this objection seems misplaced to me. Surely, believing that one will win could be a reason to think that one will. This is not always true, of course. Someone may have no chance of winning a physical contest, say, because he's vastly undersized and not nearly as strong and quick as he needs to be in order to have a decent chance of succeeding. If this person sincerely believed he was going to win he would be decieving himself. Nevertheless, all things being equal, believing that one will win makes it more likely that one will win. More generally, in as much as one does have a chance to succeed, the belief that one will win only makes success more likely. And so believing that one will succeed does bear, at least to some degree, on the question of whether one will.

So if one believes that *p*, one is answerable to certain questions that could be answered by the considerations that bear on whether *p*. Likewise for intention. If one intends to Φ one is answerable in the same sense. If, for example, someone asked me why I decided to go to the grocery store now rather than later, I could respond by stating my reasons for going—reasons that bear on whether to Φ. My intention to go to the grocery store now is something that I am committed to by virtue of reasons I have for going.

These need not be good reasons. It might be the case that my reasons for going to the grocery store now are based on misinformation. Thus, the considerations that bear on whether to go to the grocery store now do not in fact justify my intention to go. Nonetheless, I take them to be good reasons, and that's all that's important here.[43] My attitude might be amended (hopefully) if someone were to point out to me that my reasons for going to the grocery store now are not good ones. But the point is that I am committed to my intention by considerations that I take to bear on the question of whether to go to the grocery store now or later. They may not in fact bear on this question, but so long as I take them to, I can appeal to them if confronted with questions and criticisms about whether to go.

Thus, Hieronymi's solution to the wrong kinds of reasons problem: the right kinds of reasons are those that bear on a certain question, the settling of which constitutes forming a certain attitude. One is committed to this attitude in the sense of being answerable to questions and criticisms for having it, and one has it in as much one takes oneself to have reasons that settle the relevant question. Any reason that bears only on a different question is not of the right kind. And any attitude not commitment-constituted is not supported by reasons of the right kind.

43 The distinction between good reasons and operative reasons, or reasons that one merely takes to be good, is very important, however. As I argue below, Hieronymi does not adequately appreciate this distinction.

However, there are a few problems with this account. First, though it makes the content/ attitude distinction exclusive, this is not sufficient to show that an attitude is fitting. Reasons of the right kind will show an attitude to be fitting only if they are good ones. Return again to the example of Sally. Sally believes that there's a black snake in her closet, but suppose she's wrong. Suppose that the reason she thinks there's a black snake in her closet is that she thinks she sees one. We can even say that she's subjectively justified in her belief, given certain conditions. But even in this case, Sally's fear is not fitting. There's nothing in fact for her to be afraid of. Like believing something on the basis of a premise that isn't true, there's a difference between settling a question in one's own mind and actually settling it. For example, say I'm wondering whether the mail came today. Suppose that, instead of actually going out to the mail box and checking to see if the mail came today, I simply look at the time and assume that it came. In this case, I jump to the conclusion that the mail came today without knowing for sure that it did. But suppose I think the matter settled anyway. I have settled, in my own mind, the question of whether the mail came today, but the question is not in fact settled by the evidence I take to settle it.

The account does not appreciate this distinction. Consider: "constitutive reasons *bear or are taken to bear* on a question, the answering of which amounts to having the attitude...[but] there is room for various kinds of mistake: mistake of fact, mistake about whether the consideration bears on the question, and mistake about whether a consideration that bears on a question settles the question" (2005:449, original italics). Hieronymi grants that reasons that are merely taken to bear on the question of whether *p* do not necessarily settle the question of whether *p*, but does not appear to recognize that this could make the relevant attitude unfitting.

We may think our feelings appropriate because we bear in mind certain considerations that seem to us to settle the question of whether some object is Φ. But unless the object of our emotion really is Φ, our feeling is not fitting.

Second, the account is not sufficiently generalizable. Hieronymi's solution is relevant to any attitude that is constituted by the settling of a question. The question, then, is, are the emotions commitment-constituted attitudes? Though she does not specifically address this question. Certain remarks she makes suggest her answer is 'yes.'[44] Consider:

> in order to determine the right kind of reasons for the vast and interesting range of attitudes with which the current discussion is concerned (attitudes such as admiring, preferring, fearing, envying, valuing, and desiring), one will have to determine the commitment(s) implicit in each. That is, one will have to identify the question or set of questions, the settling of which amounts to having the attitude.... The wrong kinds of reasons problem, itself, suggests that there is some such distinction [between constitutive and extrinsic reasons for it]. Thus, it seems likely that, insofar as the problem can arise for these attitude, they are, in fact, commitment-constituted. (2005:454-455)

The conclusion is expressed weakly, with the phrase "it seems likely that," but the fact is that the wrong kinds of reasons problem can arise for attitudes that are not commitment-constituted, and the emotions are one such kind. To see this, notice that there are two sides to a commitment-constituted attitude. We can distinguish between the right and wrong kinds of

44 Hieronymi has admitted in conversation that her account does not appear generalizable to the emotions. However, she holds out hope that it will one day be able to accommodate them.

reasons for having certain attitudes. Such attitudes are amenable to rational criticism in the sense that we can ask whether one is justified in one's attitude, where this means something like 'is the attitude fitting?' We can say that one is committed to one's attitude insofar as one is answerable to this kind of question. And to defend one's attitude in this way is just to give reasons that bear on the right question. This is the commitment side of a commitment-constituted attitude. The constituted side is this: that the bearing in mind of reasons one takes to settle the relevant question amounts to forming some attitude.

The settling of a question makes one committed to the relevant attitude. If Sally bears in mind certain considerations that would settle the question of whether *p*, we can say that she is committed to an attitude by virtue of these considerations. It just so happens that the settling of this question constitutes believing that *p*. So she is not committed to this belief just in virtue of having it. Rather, she is committed to it rationally, because of the reasons she has that *prima facie* settle the question of whether *p*, i.e., because she is answerable to certain questions and criticisms that call on her to justify her belief.

The emotions are commitment-type attitudes. Justifying the way one feels is similar to bearing in mind reasons that settle one's belief, since these two exercises aim to settle the same question, whether *p*. Whether Sally's emotion is fitting is answerable by reasons she has to believe that the object of one's emotion is Φ. In order for shame to be fitting the object of one's shame must be shameful. One's fear is rationally assessable by whether the object of her fear is dangerous, etc. Thus, Mary could very well defend her feeling of shame by citing reasons which bear on whether the object of her shame is shameful. She can cite aspects of the object of her

shame—that her job was really not that hard, that she should have been able to do it, that she failed anyway—that bear on whether whether her failure was shameful.

However, emotions are not constituted by attending to reasons that one takes to settle the question of whether *p*. Return to the example of the disaffected sloth. He judges his state shameful but feels no corresponding shame. He's not even disposed to feel shame. The settling of the question of whether *p* does not, when it comes to his feelings, necessarily constitute having some feeling. Emotion and judgment can notoriously come apart.

Third, the account produces a result similar to Parfit's and Gibbard's. Recall how both construe the matter: attitude-given reasons are not 'really' reasons for having that attitude, but are reasons for wanting to have that attitude. By showing that an attitude would be good to have, one does not thereby justify having it, but only justifies one's desire to have it. Hieronymi differentiates right reasons from wrong by distinguishing different questions. For belief, this is the question of whether *p* and whether the belief that *p* would be good to have. Any reason that does not bear on settling this question is not really a reason to have that belief. It is a reason to have a belief about that belief (that it is good to have). As she puts it, "extrinsic reasons are not 'really' reasons for believing *p*, we can say, because they are not the kind of reasons which, simply by finding convincing, one would believe *p*" (2005:448). Hieronymi's solution to the wrong kinds of reasons problem thus is also a way of making it disappear.

But it seems odd to say that there are no reasons pertaining to the beneficial upshot of believing *p* that justify believing *p*. Suppose there are many ways in which the belief that god exists would be good for one to have, but that one has little evidence for so believing. Must we also say that there is no reason for this person to believe that god exists? It would be

intellectually dishonest, perhaps, but that doesn't mean that one's belief couldn't be justified all things considered. If this person were to make her life better by so believing, there seems to be a straightforward sense in which her belief is rational. This point can also be made for the emotions. The fact that an emotion is good to have seems like a perfectly good reason to have it. Though it's important to be able to distinguish those reasons which show whether an emotion fitting from those that don't, it does not follow that reasons which do not bear on this question play no part in justifying that attitude.

These concerns suggest a few guidelines for a solution to the wrong kinds of reasons problem. First, the solution must work for the emotions. We need to make sense out of how an emotion is justified in the right way. An account that does not address this concern will not be of use. Second, a solution should not deflate the justificatory status of moral or prudential reasons for feeling F. These other kinds of considerations are potentially justifying. That is part and parcel of why the wrong kinds of reasons problem is a problem. Different kinds of reasons may justify an emotion, but only certain kinds do it in the right way. The question is how we are to understand the right kinds from the wrong kinds when both are potentially justifying. Finally, a solution to the wrong kinds of reasons problem must be sensitive to the necessity that the right kinds of reasons are good reasons: an emotion is not fitting if it's object is not really Φ. An emotion's fit bears on the properties of an object. Without the relevant properties, we don't get fittingness. The following section details a solution to the wrong kinds of reasons that meets these desiderata.

1.4.3 The Dual Role of Reasons

Rabinowicz and Ronnow-Rasmussen have recently offered another solution to the wrong kinds of reasons problem that works for the emotions. They propose that the right kinds of reasons have a dual role. On the one hand "(i) they appear in the intentional content of favoring as the features on account of which the object is favored. On the other hand, (ii) they justify favoring the object in that way, that is, provide reasons for the pro-attitude in question" (2004:414). This way of construing the matter is meant to capture the insight that our attitudes are fitting if they are held on account of some properties of an object. Fitting attitudes are discerning, and the features of the object to which they are fit justify feeling them. The right kinds of reasons appear to have a dual role.

The authors give an example of cherishing unspoiled wilderness (2004:414). This wilderness is valuable on account of its being unspoiled. This property of this wilderness has a dual role. It fills out the content of our cherishing and justifies our attitude towards it. We can say that our cherishing of such a place is directed towards or about this property of the wilderness, and that our attitude is fitting because the wilderness is unspoiled. Or consider an example using the emotions. Say that Sally encounters a black poisonous snake in her closet and is afraid. The ability of the snake to harm Sally both justifies her fear and is what her fear is about. On this view, Sally's fear is fitting because her reason for being afraid appears in the content of her emotion and justifies her feeling it.

The wrong kinds of reasons do not fulfill this dual role. They could fail to justify their attitudes in one of two ways. First, the object of one's attitude may not really have the properties

that would justify having the attitude. Suppose that, instead of a poisonous black snake, Sally encounters a harmless garden snake and is afraid. In this case, the content of her fear is directed towards something she takes to be dangerous. But the property she takes the snake to have—presumably it's being poisonous—the garden snake lacks. Thus, the reason which appears in the content of her attitude does not justify it. It is the wrong kind of reason.

Second, reasons can fail to justify an attitude in the right way if they do not speak to the characteristic concern of that attitude. The authors give an example that makes use of the paradox of hedonism.

> On the hedonist view, we are supposed to care for our friend for his own sake, because *caring for him for his own sake would make us happier.* The italicized property of the person in question is a reason for caring, but are we supposed to care for him on account of that reason? Surely not. Caring for him on account of that property would not make us happier. To promote our happiness we should care for our friend for his kindness, humor, intelligence, the things we have experienced together, and so on...That caring for him would make us happier is a reason for caring but it is not what should be our motive for caring. It is not the feature that is supposed to occur in the intentional content of caring as the ground for that attitude (Rabinowicz and Ronnow-Rasmussen 2004:415).

Care is a discerning attitude. It is an attitude directed towards those features of an object in virtue of which it is valuable. We could care for our friends because that would make us happier. But to have this as the object of our concern suggests something perverse about our attitude. To

care for someone is to value, respect, and have strong positive feeling for them in virtue of that which makes them good friends. The attitude is characteristically concerned with those features.

Some attitudes appear to be nondiscerning in the sense that there appears in their intentional content no features which would justify having them. The familiar romantic version of unconditional love, for example, seems to be such an attitude. One loves one's spouse in this way not in virtue of any feature he has, but unconditionally. It would seem, then, that such attitudes could never be fitting. And since RDT expresses a biconditional relation between fittingness and value, the objects of unconditional love could not be valuable either on this account. Needless to say, this seems dissatisfying.[45]

But this may not be so problematic. Consider Harry and Sally. Harry loves Sally unconditionally; he loves her regardless of whatever qualities she has. Even if she were to change dramatically, he would still love her. Though this seems endearing, such an attitude can go too far. Imagine that Sally has no admirable qualities. Say she's despicable in every respect. Is Harry's love fitting? It seems like we would want a friend to give him some good advice and leave her, and for the relevant reasons—that there's nothing about Sally to love. Or say that Sally starts having an affair with another man and rejects Harry. Harry's love for Sally is misplaced. We can even imagine Sally changing to such a degree that it would make sense for Harry to find someone else to be the object of his affections—if Sally were to go into a coma, for instance, or otherwise become somehow unresponsive to Harry—or she suffered from some sort of brain disorder and started acting completely unlike herself, etc. Furthermore, as the authors point out, even if it is required that one love one's spouse unconditionally, this doesn't

45 Unconditional love is the only attitude of this sort I can think of. Rabinowicz and Ronnow-Rasmussen also mention this in their discussion, but no other (2004:416).

necessarily threaten the value of one's spouse. So long as he or she does have qualities that would justify a discerning kind of affection, one could love unconditionally. It would be perfectly okay, for instance, if Harry loved Sally unconditionally so long as she also had qualities towards which a discerning love would be fitting. Undiscerning attitudes don't necessarily imply that their objects are not valuable (2004:416).

Some discerning attitudes appear to violate the dual-role rule. Recall an earlier case, about the evil demon who demands that we admire him for his own sake lest we be punished severely (Rabinowicz and Ronnow-Rasmussen, 2004:419). This example threatens the account if it is to apply to attitudes that do not have a characteristic concern. We are to admire the demon in this case because of his determination to punish us if we don't. But we are to admire him just for his determination, that is, for his own sake. It seems here that admiring the demon for this reason satisfies the dual-role criteria. The demon's determination to punish us if we don't admire him justifies our admiration, and it appears in the content of our admiration. Our admiration would be based on the right kind of reason, on this account, but it doesn't seem that the demon is admirable. This appears to be a counterexample to the proposal.

However, to admire someone is to be impressed by some strength or achievement on her part. I admire Mr. Rogers because of his ability to improve the lives of many children. So too do we admire the demon because of a strength, namely, his determination to punish us if we don't admire him. But the reason that justifies our admiration is not his determination to punish us; it's the consequences of not admiring him. Thus, what justifies our admiration is at odds with the characteristic concern of our admiration. As the authors put it, there is a "disparity between the way in which the property justifies the attitude and the characteristic concerns of the attitude

in question" (2004:420-421). The property of an object that justifies one's response towards it must not only appear in the content of one's response, it must justify that response in the way that response is characteristically concerned. The demon's determination to punish us does not justify our admiration in this way, and so admiring him on this score is not admiring him for the right reason.

Thus, the dual role of reasons account must be restricted to attitudes that have a characteristic concern. It will not work, for instance, with attitudes such as desires and preferences. I may desire all kinds of things none of which have any kind of property that speaks to the nature of my desire. I could desire, for instance, ice cream, to take a walk, to sleep late, to speak out in favor of Mr. Roger's Neighborhood, etc. If there is something these objects have in common that speaks to my desiring I don't know what it is. On the other hand, the account works very well with the emotions. The emotions can be individuated according to their characteristic concerns, so it's quite easy to isolate those reasons that will speak to their fit. Rabinowicz and Ronnow-Rasmussen see this as a weakness of their solution, or rather, since their solution is offered for use in any response-dependent analysis, a weakness of the analysis. Be that as it may, this isn't a problem for the sentimentalist. So long as value is analyzed in terms of thick attitudes, such as the emotions, the dual role of reasons account provides a way of understanding the difference between reasons of the right kind and reasons of the wrong kind.

Consider now two implications of the dual role of reasons account. First, the second role of reasons, that they justify the attitude in question, should be read weakly. It is not necessary that the properties that figure into justifying the attitude do so in an overriding way. Rather, they only need provide some justification for that attitude. Thus, reasons of the wrong kind can

outweigh reasons of the right kind, but they do not show reasons of the right kind to be spurious. RDT posits a biconditional relation between fittingness and value. So not only should reasons to have some attitude imply value, the reverse should also be true. But this says nothing to the effect that the right kinds of reasons are incapable of being overridden by the wrong kinds. If Mary's failure at work is really shameful, it must be the case that there are certain properties of her failure that figure into the content of her shame and justify feeling it. There may be very good reasons not to be ashamed—if it would prevent her from looking for more work, for example. And these other kinds of reasons could very well offer more weight than the fitness ones. We needn't discredit the wrong kinds of reasons as somehow not 'really' reasons to have some attitude.

Finally, we should anticipate the possibility that some properties of an object could stand in favor of justifying an attitude while others stand against it. An object can be valuable in some respects and not others, and we should expect to encounter a problem of aggregation on some occasions, perhaps often. There may be some respects in virtue of which Mary's failure is fitting but perhaps others in which it isn't. Say she suffered from illness much of the time she was trying to adjust to her new job. Despite the fact that she was well trained for it, and that it wasn't that hard, Mary had to overcome at least this obstacle. This counts against thinking her shame fitting. More generally, we can say that whether an object is valuable will depend on an assessment of the right kinds of reasons for having some response. Resolving a balance between conflicting reasons is a difficult and complicated matter, and I have nothing to say about how to do that here.[46] The point is that only the right kinds of reasons will count in determining this

46 For some reflections on this issue, see Rabinowicz and Ronnow-Rasmussen (2004:418, n. 76).

balance. The wrong kinds of reasons may speak to whether Mary should feel shame, but not to whether her shame is fitting.

The dual role of reasons account meets the desiderata for making the distinction between the right kinds of reasons and the wrong kinds of reasons. It works for the emotions. It allows that the wrong kinds of reasons are 'really' reasons that can count in favor of holding an attitude and it is sensitive to the fact that the object of one's attitude actually be Φ if it is to justify having that attitude.

Further, the fact that the solution only works for discerning attitudes provides an upshot for a moral version of rational sentimentalism. On this view, recall, the right kinds of reasons are those that speak to the fittingness of guilt. As guilt is characteristically concerned with wrongdoing, the right kind of reason for feeling guilt is the properties of Φ-ing that constitute it as morally wrong.

This, however, raises a further difficulty. If claims about wrongdoing are analyzed into claims about the fittingness of guilt, and claims about the fittingness of guilt is understood in terms of an action being wrong to do, then the analysis appears circular. It doesn't seem to get us anywhere other than where we started. In the next chapter, I show that the threat of circularity can be avoided. Guilt, while characteristically concerned with wrongdoing, also has a distinctive feel, which can be referenced by claims about its fit. When we say that guilt is fitting, then, we don't say that the feeling concerned with moral transgression is fitting, but that 'this feeling' (whatever it is that constitutes guilt) is fitting. This ties moral judgment essentially to the feeling state of guilt, instead of to the content of the moral concept WRONG. And this is how it should be. Moral concepts are not substantively fixed. The fact that moral concepts are essentially

contestable encourages the thought that moral dispute is not so much about what to think, but about how to feel.

Chapter Two

Judgment and Response

2.1 Essential Contestability and Univocity

Rational sentimentalism purports to accommodate two core intuitions about moral judgment. On the one hand, moral matters appear to be essentially contestable. Two people can agree to all of the facts about a situation and yet disagree about the status of its value. This condition reflects different standards that disputants may bring to bear on their judgments about right and wrong. On the other hand, moral argument does not appear to be a kind of case where disparate parties simply talk past one another. Despite their different views about morality and what counts as an instance of moral wrongdoing, opposing sides appear to operate from a fixed point of dispute. These two qualities of moral judgment have been called the essential contestability of moral concepts and the univocity of moral dispute. Sentimentalism can claim to accommodate these aspects by tying moral meaning to the fittingness of certain responses such as guilt. The type of action or practice referred to in a moral claim gains its moral status by the attitude directed towards it.

The intractability of moral argument does not appear to be due to our inability to assess what the moral facts are. That is to say, conflict in such disputes does not appear to be merely psychological. It's not an issue that has anything necessarily to do with what we are or are not

capable of perceiving or what psychological and sociological influences we bear. Rather, it suggests something of metaethical import. It is possible, in principle, to agree about all of the natural facts and yet disagree about whether Φ-ing is wrong to do.

Failure to make headway in moral dispute may often derive from the use of two different concepts of WRONG. Return again to the likely debate between deontologist and consequentialist. Suppose killing an innocent person would produce the greatest amount of nonmoral goodness. Both will agree to the facts of the case and yet disagree over whether the act in question is wrong. Their controversy is, in part, over the extension of the concept WRONG, but is one in which each party employs their own concept. Thus, there's a sense in which they're talking past each other. One says the action is wrong because it involves the killing of an innocent person. The other says it is right because it produces the greatest amount of nonmoral goodness. Yet each disagrees with the reasons the other offers for their view. Despite recognition of the same set of natural facts, they disagree over the moral status of the action. Agreement over what's to count as wrong is also likely to be just as intractable since their situation is akin to having different intuitions about the matter. Not only do the parties disagree about the application of the concept WRONG, they disagree about what counts as 'wrong.'[47]

We can also disagree about the extension of a shared concept. In cases of this sort disagreement stems from controversy over what the natural facts are. Were opposing parties to get clear on these facts they would resolve their dispute. For example, I might agree with Ann that killing an innocent person is morally wrong to do but disagree with her about whether the destruction of a fetus is an instance of this injustice. Our disagreement turns on whether the fetus counts as a person and if we could come to agreement on this point we would be able to

47 See D'Arms (2006) for a helpful discussion of this idea. It also appears in Wiggins (1987).

resolve our dispute (presuming we are both rational). Here, we quarrel over the extension of a shared concept.[48]

To say that a concept is essentially contestable is not to say that it is vague. No doubt moral concepts are vague at least to the extent that they admit of borderline cases. If the killing of an innocent person, for example, partially fills out our concept WRONG, we can expect there to be times when we cannot say whether or not the destruction of an animal (human or otherwise) would be a moral violation. Neither does the notion suggest that one cannot make an error in application of a concept. If disputes over whether Φ-ing is an instance of wrongdoing can be settled by appeal to the natural facts of the case, then we would expect these disputes to be resolvable most of the time. But one could err in one's application of the concept. Imagine that both Ann and I agree that the killing of an innocent person is wrong and that the fetus at some stage in its development is a person. But Ann concludes that killing a fetus at this stage in its development is wrong and I don't. The intractability of this dispute is due to my failure to draw the conclusion that clearly follows in this case, not from any essential play in the conceptual filling of WRONG.[49]

48 It might turn out that the natural facts of the matter don't settle whether Φ-ing falls under the extension of this concept. There may be, for instance, no specific point at which we can say whether the fetus is a 'person,' where borderline cases present the impossibility of saying whether or not the fetus could be awarded personhood status. The concept PERSON, in other words, could be vague.

49 None of this implies that psychological explanations for the intractability of moral dispute cannot and are not often good ones. There are psychological and sociological factors that give rise to stubbornness in moral thinking. Religious conviction also appears to have been and continues to be a bad influence on moral thinking. Perhaps not so much now as in the past, religious institutions have directly influenced how we think and behave, taking on the role of censors and sanctioners. Moreover, taking a position on faith fundamentally impedes moral thinking; it limits what's to count as a moral possibility. Argument over whether abortion is morally wrong, for example, may be trumped by the belief that the conclusion has already been settled. And when religiously motivated parties argue for or against certain policies, they may regard their premises as unassailable. As Brink forcefully notes,

> moral thinking...is subject to various distorting influences such as particular conceptions of self-interest, prejudice, and other forms of social ideology. Because the subject matter of ethics concerns, among other things, the appropriate distribution of the benefits and burdens of social and personal interaction, these distorting influences often afflict moral thinking more than scientific thinking; it is

Thus, a concept is essentially contestable if one can use it without linguistic impropriety and with full awareness of the non-moral facts without (logically) having to agree with another who does the same. That moral concepts have this property speaks to a gap between the natural facts of a case and its moral status, something that many have shown great appreciation for. Hume (1739/1978), for instance, famously asserted that one cannot derive an 'ought' from an 'is'; that no normative claim follows from a consistent set of non-normative premises. Later, Moore (1903) proclaimed that the concept GOOD cannot be analyzed in terms of some set of natural properties, since it is always an open question whether that set is 'good.'[50] Though arriving at very different conclusions about what our moral terms signify, both thinkers were impressed with the apparent gap between what is the case and what ought to be. One who endorses the essential contestability of moral claims expresses the same appreciation.

The essential contestability of moral concepts allows for moral dispute without agreement on a concept of WRONG. Since neither party in such disputes can be criticized for holding the standard she brings to bear, neither party is linguistically confused. And this, in effect, separates moral dispute in these cases from similar kinds of nonmoral dispute. Suppose, for example, I think a bear is just a large dog. My friends are puzzled when I comment on how well their 'bears' are behaved. 'Fido is not a bear,' one remarks, 'he's a dog.' But this strikes me as strange since I believe that bears are just large dogs. A dispute over whether Fido is a bear would be disconnected at its core due to the fact that my friend and I have different concepts of BEAR. But unlike the moral case one of the involved parties here is linguistically confused. I have the wrong concept of BEAR and so disagree with my friend about its extension. In the moral

just such issues on which these distorting mechanisms are most likely to operate...And these sources of distortion are hardy perennials (1989:205).

50 Moore's concern in the *Principia* is over identity, not entailment. See Brink (1989:Chapter 6) and Blackburn (1973).

case neither party is linguistically confused. Consequentialist and deontologist have different concepts of WRONG but the essential play in its meaning allows for this divergence without semantical error. Despite wielding different concepts of WRONG, they are moored in a fixed point of dispute.

Rational sentimentalism proposes to make good sense out of what this fixed point is. We can say that even when two parties disagree about what makes Φ-ing wrong to do and therefore disagree about whether Φ-ing is wrong—even when this is the case—the point of connection in the dispute is how to regard Φ-ing. One thinks it a proper extension of the concept WRONG, the other disagrees. One condones it, the other condemns it. Their reasons for their particular view may cause the disagreement to diverge substantially since in this case there isn't one concept of WRONG at play here. But they converge on how this case is to be regarded morally. That is to say, they are arguing about whether guilt is a fitting response to Φ -ing. What moors their discussion is thus not the concept WRONG but the concept GUILT. One says that Φ-ing was wrong while the other disagrees, and no point of argument could persuade the other to change his mind since each understands 'wrong' in a way the other does not. However, if both had the same concept of GUILT, then they could disagree about what counts as an instance of wrongdoing in a meaningful way.

At first glance this appears problematic. Emotions are construed as having some content that can be filled out by incorporating a propositional gloss or 'characteristic concern.' Characteristic concerns are very useful tools for type-identifying the emotions and assessing their rationality. Consider, for example, the characteristic concern of guilt, *having transgressed a moral imperative* (Lazarus 1991:122). This gloss allows us to say that guilt is just that self-

directed feeling directed towards any instance of moral transgression. In addition, if the object towards which it is directed is not such an instance, then one's guilt in that case is not fitting. The problem with type-identifying guilt by its characteristic concern is that the gloss we use to describe guilt invokes the concepts MORAL and IMPERATIVE. If these concepts are tied to the meaning of WRONG deployed in an agent's evaluative judgments, then the content of one's concept of guilt will vary with the content of one's moral beliefs. Thus, people with different concepts of WRONG will not have a concept of a shared response which might ground their normative dispute. If sentimentalism is to explain the essential contestability of moral concepts alongside the univocity of moral dispute, it must explain how guilt could have a propositional gloss—one that includes the concept WRONG—without inheriting the ambiguity of the concept.

A solution to this difficulty precludes the possibility of describing emotions as themselves judgments. Emotions have propositional contents but they are not themselves judgments. Rather, they are not simply judgments. They also have an affective feel. The phenomenological aspect of emotional experience affords a way to type identify the emotions without appeal to a propositional gloss. Thus, we can talk about guilt just as a feeling, by a description of what it is like to experience a particular kind of physical state. Of course, we can also analyze guilt by using moral concepts. Indeed, construing the emotion in this way is helpful for several reasons.[51] It just isn't necessary that we use the concept WRONG to analyze the concept GUILT. In this sense, the characteristic concern of guilt is tied to a concept of wrongdoing, but its affective part is not. The separate affective aspect of guilt allows us to explain how two people can bring different moral standards to bear in an argument and still be united in a common point of reference. One thinks that guilt is fitting and the other does not. But what each means by

51 See below.

'guilt' need not inherit the ambiguity of moral wrongdoing. Rather, it is as if one says that 'one should feel this way towards Φ-ing' while the other disagrees. And what each means by 'feeling this way' is the same for each.

2.2 Emotions and Judgments

What is the difference between an emotion and a pain? Both have a feel aspect. Guilt feels bad and so does pain. Physical pain is usually isolated to one area of the body, guilt isn't. Emotional pain may be more like guilt in this respect. It's even tempting to think of guilt as a type of emotional pain. One important difference between emotions and pains, however, is that emotions seem to be about something. One's anger has an intentional content; one's pain does not. The discomfort in my foot is not about anything at all, but is simply there. It may have some cause, but it is not a mental state with any content. My guilt, on the other hand, is about something. If I steal a cookie from the cookie jar and feel guilty, my guilt is directed towards the wrongdoing that I unfortunately committed.

Another difference between emotions and pains is that emotions can be irrational while pains cannot. If I stub my toe and it starts to throb I can't, nor can anyone else, criticize the throbbing in my toe for somehow not making any sense. Sometimes pains are felt mysteriously, such as 'phantom' pains felt absent the relevant part of the body (e.g., pain in a missing leg). These pains are mysterious in the sense that the presence is surprising and an explanation for them hard to come by.[52] Still, they are not irrational in the sense that one has somehow made an error in feeling it. Whatever conditions have conspired to give rise to pain, they can serve as an

52 Damasio (1994) has recently given an account of phantom pains. See also Prinz (2004a:6).

explanation of its occurrence. But there are no normative constraints on whether one is to feel pain. The same is not true for the emotions. Aside from the moral and prudential upshot of some feelings, emotions can be fitting or unfitting. It does not make sense to feel sad at having won the lottery (all things being equal), or to feel guilty for doing something that is clearly in the right.

A convenient way of explaining these two features of emotion—their fittingness and their intentionality—is to introduce the notion of characteristic concern. An emotion's characteristic concern is what it is generally about, though this is a loose way of putting it. It's loose because the terms 'about' or 'directed towards' in this context are ambiguous. An emotion can be directed towards or about a particular event such as when I feel guilty for taking a cookie from the cookie jar. And emotions can be directed towards or about an object in virtue of which it is directed towards a particular event such as the fact that my stealing from the cookie jar is wrong. In the latter sense, my guilt is about a moral transgression. Call this distinction that between an emotion's 'formal object' and it's 'particular object.'[53] An emotion's formal object is the property in virtue of which an event elicits an emotion while its particular object is the event itself. Thus, my guilt is elicited by the particular event of my stealing from the cookie jar because stealing from the cookie jar is something I take to be a moral transgression. An emotion's characteristic concern is directed towards its formal object, not its particular one.[54]

53 This terminology derives from Kenny (1963). See also Prinz (2004), and Greenspan (1988).

54 Some mental states are distinguished from emotions precisely because they are not about anything in particular. Free-floating anxiety, or depression—moods, in other words—are not about any event in particular. According to the prevailing view, emotions are. If I feel guilty I feel guilty about something in particular that I did, even if I think that I shouldn't. But emotions are not characteristically concerned with the particular objects that they are about. My guilt is not characteristically concerned with my stealing from the cookie jar, but with what I take to be a moral transgression.

Putting the matter in this way is helpful for a few reasons. First, it allows us to make sense of how emotions are rationally assessable. For example, say I feel guilty for stealing a cookie from the cookie jar. Suppose further that such a thing is morally wrong to do. If it is, then my guilt is fitting. By contrast, it would not be fitting for me to feel sadness after stealing from the cookie jar since sadness is characteristically concerned with incurring some loss. What allows us to say that sadness is unfitting is the dissonance between its formal and particular objects. Sadness is not fitting because stealing a cookie from the cookie jar is not an instance of loss. Or take this example, while it may be fitting to feel guilty for stealing a cookie from the cookie jar it would not make sense to feel guilty for helping an elderly lady cross the street. I may regard this event as an instance of moral wrongdoing but I would be mistaken about that (presumably). The particular object of my feeling (helping an elderly person) is at odds with its formal object (moral transgression).

Second, we can individuate the emotions according to what they are characteristically concerned with. Consider some suggestions recently proposed by Lazarus (1991:122). Anger is identified by its characteristic concern *a demeaning offense against me and mine*, guilt with *having transgressed a moral imperative*, shame with *having failed to live up to an ego-ideal*, etc. These glosses provide an intuitive way of parsing the emotions out. Notice also that appeal to an emotion's particular object does not do nearly so well in this regard. Two people might feel guilty for different things but would not feel different emotions on that account. For example, say that I feel guilty for stealing from the cookie jar while you feel guilty for embezzling money from the bank. Our feelings have different particular objects but are not different feelings. It is the formal object that they share, *moral transgression*, which allows us to say that we both

experience the same feeling. Or suppose that two people experience different emotions towards the same particular object. Say that I feel guilty at having taken a cookie from the cookie jar while you feel pride (for having taken one yourself). Suppose further that you regard your action as a success worthy of great respect. The fact that our feelings have the same particular object does not say anything about whether they are the same feeling. What differentiates your feeling and mine is their formal objects. I regard my deed as a moral transgression while you construe yours as a successful endeavor.

Third, type-identifying the emotions according to their formal objects allows us to maintain that the corresponding evaluative concepts are essentially contestable. Deontologist and consequentialist disagree about whether killing an innocent person is wrong in some cases. The natural facts of the situation do not determine how to judge in this situation. The deontologist thinks that the violation of a moral rule constitutes wrong action while the consequentialist believes that wrongness consists in some feature of the consequences that result from Φ-ing—and neither appear able to unseat the intuitions of the other. Thus, the concept WRONG does not necessarily refer to some actions or kinds of actions and not others. Rather, its extension appears to depend upon what kinds of actions we take to be wrong. Likewise, guilt is not characteristically concerned, necessarily, with some actions or kinds of actions and not others but rather with whatever one takes to be a moral infraction. The play in the concept WRONG also exists in the formal object of guilt.

In all cases, emotions depict their objects as having value. Indeed, their characteristic concerns seem to express the object's significant value. As Nussbaum remarks, "fear requires the thought that important damages can happen to us through no fault of our own; anger, again,

requires the thought that the item slighted by another is of serious value. I do not go around fearing that my coffee cup will break; I am not angry if someone takes a paper clip. I do not pity someone who has lost a toothbrush. My breakfast cereal does not fill me with joy and delight" (1994:370). Roberts, too, describes emotions as 'concern based,' where this means having something to do with those things we take seriously. As he says, "to be angry is not just to see a person as having culpably offended; it requires a concern about some dimension of the offense, and possibly a concern about some dimension of the offender. To be afraid of heights is not just to see them as a danger to something-or-other; it requires that something I hold dear seem threatened" (1988:191).

Thus, characteristic concerns are those standards that express the high value of an object against which we can measure an emotion's fit. We can use these standards to individuate the emotions and to assess their rationality (fittingness). Mary's shame is fitting only if the particular object of her shame, having failed at her job, is something that really matters. My fear of the black snake in my closet is only fitting if it is in fact dangerous. The value that any emotion's characteristic concern expresses must be construed independently of one's disposition to feel that emotion—lest we encounter a vicious circularity. We can't construe the value that fear is supposed to pick out as simply those things that make us afraid. The value that guilt is supposed to be responsive to isn't just that which makes us feel guilty. An emotion's characteristic concern is a standard against which we can measure an emotion's fit. If we refer only to those things which reliably elicit our feelings we don't leave any space to account for the sense in which we are supposed to feel them.

The natural inclination is thus to presume that emotions themselves are judgments. Characteristic concerns are in propositional form, making it natural to think that they're contents of judgments. Take guilt. Guilt's characteristic concern is *moral transgression.* The idea would be that to feel guilty is to somehow judge or assess the particular object of one's emotion, the particular act that one has performed, as an instance of wrongdoing. It is to make an evaluative judgment of that emotion's particular object. Further, we can appeal to the content of that judgment to explain why it amounts to guilt rather than some other reaction such as a form of anxiety or some other feeling state. We can also use the propositional content of guilt to explain why it would amount to a fitting response to its particular object—stealing from the cookie jar, say. If emotions are judgments, their characteristic concerns would be just the contents of those judgments.

One way to interpret this idea is to hold that emotions themselves are or are necessarily caused by judgments or cognitive mental states of some sort.[55] On this account, normally called 'judgmentalism,' when I feel F I'm actually judging or appraising my environment as somehow having some evaluative feature, Φ.[56] On a different interpretation, it is essential that, in order to type-identify and rationally assess our feelings, we must employ propositional glosses to serve as their characteristic concerns. This view does not assert that emotions are judgments *per se* but rather that judgments serve a methodological purpose. This approach has been aptly been called 'quasijudgmentalism.'

55 For simplicity's sake, I refer only to the constitutive thesis, not the causal one.

56 Cognitivists about emotions tend to hold that evaluative judgment is necessary but not sufficient for its affective counterpart. Their view is that one can believe Φ without feeling F, but not the other way around. However, when it comes to guilt the matter becomes more controversial. Internalists about motivation, for instance, may think it obvious that believing that one has done something wrong entails that one feel guilt. I explore some internalisms about motivation in Chapter Three.

Both views pose problems for sentimentalism. Many have noticed, for example, that a cognitivist reading of the emotions makes any sentimentalism circular. If guilt involves the judgment that one has violated some moral obligation, then analyzing moral judgment in terms of guilt simply takes us back to where we started. The very judgment we seek to analyze is contained in the emotion used to elucidate its meaning.

Another related concern is sentimentalism's ability to reconcile the essential contestability of moral concepts alongside the univocity of moral dispute. Judgmentalism implies that the judgment involved in emotion employs concepts. These concepts appear in the that-clause which ascribes the belief, such as in the belief *that it is raining*, or the belief *that stealing from the cookie jar is wrong*. If emotions are constituted, at least in part, by beliefs, then having an emotion would require having those concepts used to ascribe them. Thus, one who believed that stealing from the cookie jar was wrong would also have the concepts STEALING, COOKIE, JAR, and WRONG.[57] The problem for (rational) sentimentalism, then, is this. As I argued in section **2.1**, two parties can have different concepts of WRONG and yet still engage each other meaningfully in dispute over its extension. What moors their discussion is the response that they share (guilt). But if this response itself contains the concept WRONG, then disputants will not have the same response to which they can appeal. Thus, on a cognitivist reading of the emotions, sentimentalism appears to becomes unable to account for univocity of moral dispute in these cases.

There are several points in favor of judgmentalism. First, there is a close relationship between judgment and feeling. By most accounts, believing that something would be wrong to do suggests or entails having some motivation not to do it. Likewise, if one changes one's

57 Presuming that this person is not believing in an 'inverted commas' sense (see Hare 1952).

beliefs about what is morally right to do, then one's motivations will also change. So at least in this respect there is an interesting relationship between belief and motivation. And judgmentalism anticipates these phenomena, for if emotions are constituted by judgments, changing one's judgments would elicit changes in one's feelings. Emotions are motivating agents and if they were themselves judgments, the consonance we find between moral belief and motivation might be similar to the demands for consistency we place on our beliefs.

Second, some feelings necessarily involve some thought or belief. One cannot be sad that one has lost one's bicycle, for example, without believing that one has lost one's bicycle. One cannot be afraid of the monster under the bed without believing that there is a monster under the bed. These emotions might be described in terms of the specific beliefs which accompany them, such as monster-under-the-bed-fear, or bicycle-sadness. D'Arms and Jacobson describe a colorful example of what they call 'tenure-rage,' which is the anger one feels at having been denied tenure (2003). These kinds of emotions they call 'cognitive sharpenings.' There are an indefinite variety of cognitive sharpenings, as many as can be characterized by the accompanying belief. Many of them we don't really have names for, like 'tenure rage.' But others exist in our canonical repertoire (such as homesickness). Cognitive sharpenings appear to bolster the judmentalist's case. These associated thoughts are likely constitutive in feeling. If so, at least some emotions are judgments.

Third, we can often infer that, if one is in some emotional state, that one is also committed to the corresponding evaluative belief. If I feel guilty for stealing a cookie from the cookie jar, I likely also believe that stealing from the cookie jar is morally wrong to do. If I resent my older sister's attempts to frame me for stealing a cookie when I didn't, it seems

reasonable to suppose that I believe that my sister acts unjustly. Further, emotions appear to be dependent upon our beliefs in this way: a change in one's beliefs elicits a change one's feelings. If I come to believe that the black snake I see in my closet is really a pile of black socks, I am likely to be relieved. This kind of change would be expected if emotions themselves were judgments. Just as we have demands of consistency on what we believe, we would demand that our feelings too would be responsive to what we think we should feel.

However, the case for judgmentalism also faces some problems. First, the question about the cognitive status of emotions is not whether some emotions are necessarily constituted by judgments but whether all of them are. Of particular interest are the paradigmatic emotion kinds, such as anger, disgust, sadness, shame, joy, guilt, etc. There are cognitive sharpenings of each, surely, but are each constituted by judgments? There may be a specific kind of shame that involves the thought that one is uncommonly short, call this 'vertically-challenged-shame.' But does shame proper involve any sort of judgment? The judgmentalist thinks so. He thinks that shame involves a specific sort of judgment, as captured by its characteristic concern, in Lazarus' words *failure to live up to an ego ideal.* The question is whether one must believe that one has failed to live up to an ego ideal to feel shame, or more precisely, whether the judgment that one has failed in this respect constitutes feeling shame.

Second, emotions can also be recalcitrant. We can feel F without holding the corresponding evaluative belief; and we can experience a change in our evaluative beliefs without experiencing a change in our feelings. The guilty ex-catholic who eats meat on Fridays during lent feels guilty for something she does not believe is wrong to do though formerly she did. The significance of this phenomenon is not necessarily problematic for judgmentalism,

however. For, even if emotion and evaluative judgment can come apart, they usually go together, and the hypothesis that emotions are themselves evaluative judgments would explain this coincidence. Emotional recalcitrance would thus be the unfortunate situation of holding inconsistent beliefs. If I believe that eating meat on Fridays is morally acceptable but feel guilty when I do it (due to my catholic upbringing), then I simply have inconsistent beliefs, one that eating meat is morally acceptable and another that eating meat is not morally acceptable. Were it uncommon that people hold inconsistent beliefs, judgmentalism would be burdened with explaining why they do in cases of emotional recalcitrance. But it isn't uncommon that people hold inconsistent beliefs.[58]

The decisive problem with judgmentalism, however, is not that emotion and evaluative judgment can come apart but that the way in which their dissonance takes shape is not like that of conflicting beliefs (Roberts 1988, Greenspan 1988). Consider again the case of the guilty ex-catholic. Let's fill this example out a little more. Say that I feel guilty for what I am doing and that I try to 'get a hold of myself.' I tell myself that there's nothing to feel guilty about, but feel guilty anyway. What exactly is going on here? One possibility is that I have the disposition to believe that eating meat on Fridays is morally acceptable but that in this case I judge episodically that eating meat is morally wrong. If this were true, I might subconsciously believe that eating meat on Fridays during lent is acceptable, while forgetting this at the moment and believing that doing such a thing is wrong. But assuming that I recall evidence for my dispositional belief—something suggested by my attempt to 'get a hold of myself'—this doesn't appear likely. The fleeting nature of the emotion also suggests that my condition is less irrational than

58 The term 'belief' is normally construed in a dispositional sense. When someone believes that X, one is disposed to assert X on some occasions, defend X on others, etc. Judging, meanwhile, may more accurately be described as an occurrent attitude. Nonetheless, I use the terms interchangeably here.

subconsciously believing that eating meat on Fridays during lent is morally acceptable while consciously thinking the opposite is true.

Another possibility is that both of my beliefs are occurrent. In this case, I am either simultaneously making contradictory judgments or very quickly toggling between contradictory judgments. But both alternatives seem to be extremely unusual and irrational psychological states. Emotional recalcitrance, on the other hand, though not desirable, is not anything so unusual nor irrational. My guilt for eating meat on Fridays during lent doesn't make sense (i.e. isn't fitting), but I do not appear to be as irrational as I would be if I were to believe that eating meat on Fridays during lent is both right and wrong. I can imagine my friends reprimanding me, 'how can you possibly believe that? It's impossible for something to be both right and wrong.' Feeling guilty for doing something one believes is okay, on the other hand, is more forgivable. I know it's okay to eat meat on Fridays during lent, but I just can't help but feel guilty for doing it. I wouldn't if I could help it, but I can't.

It also appears that we have rational control over our emotions in ways that we do not have over our beliefs. Consider an example given by Roberts:

> I am standing on a wobbly ladder, doing something important enough to warrant the risk to my bodily well-being. My judgment regarding the danger I am in is rational. And my fear is intense enough to impede me from doing the job I am on the ladder to do; so that, other things being equal, it is rational for me to try to mitigate my fear. On the judgment theory, to mitigate my fear is to change my judgment about the situation...[But it] will be practically irrational to put the danger out of my mind; I must keep the danger

> in mind so as to avoid it as far as possible. If my judgment is rational and importantly relevant to the situation, rationality usually requires that I keep it as is. However, it is not irrational to try to mitigate my fear so as to be able better to do the task that needs to be done. (1988:198-199)

It's important that I believe that I am in a dangerous situation without experiencing the negatively valanced charge of fear, suggesting that in some cases it's in one's best interest to regulate the way one feels without changing the way one thinks. An analogous argument can be made for envy. Suppose that a rival receives a promotion that you were in competition for. The conditions of the promotion, however, are contingent on her performing at a very high level, and that no one else seems better suited for the job. It may be in your best interest to believe that she has something that would be valuable to you if you had it—for this may encourage you to do better work than she and give you a shot at the position she now enjoys. However, suppose your envy is getting in the way of your performance. The fact that she got the promotion is driving you crazy and making you distracted. It may be in your best interest to eschew your envy as much as possible, but to keep in mind the fact that she has something that you want.

While fear embodies a concern for oneself, guilt embodies a concern for others, or one's duty. So believing that one has done something morally wrong does not address something that bears on one's well being, and as such, there may be little upside to believing that Φ-ing is wrong. Still, feeling guilty could come at great cost, and if it does, one might be justified in not feeling guilty for doing something that one believes is wrong. Say I steal a cookie from the cookie jar and believe that I have done something wrong. However, I feel so guilty about it that

I can't do my homework. I decide not to feel guilty about doing what I did. Now I can do my homework and not worry about the fact that I've done something wrong (or so I believe).

A better tack for the judgmentalist is to retreat to a weaker position. She might reject the idea that emotions are judgments but insist that they can be identified by means of a propositional gloss, as indicated by their characteristic concern. Emotions are like judgments in the sense that judgments too can be definition-descriptions. However, emotions are not themselves judgments. Thus, the cognitivist might offer a methodological take on how to describe the emotions, not one that involves commitments about what the emotions are. The position has been named 'quasijudgmentalism.'[59] According to quasijudgmentalism, we need propositions to describe the emotions, but this is all—we need not think that the emotions are judgments or involve the use of concepts. Quasijudgmentalists offer metaphorical descriptions of emotions. They make no attempt to analyze them or find some ontological category to fit them in. Rather, they simply provide some convenient and illuminating way of thinking about what emotional experience is like while using propositions to characterize what the experience is about. Consider Greenspan's (1988) take on the matter. Greenspan describes emotional experience as being like entertaining a 'thought.' Thoughts, on her view, do not count as beliefs themselves. Yet thoughts are evaluative assessments of a situation, and so have a propositional content. She speaks, for example, at some length about a past experience of being in a car accident that resulted from skidding on ice. Now she feels afraid whenever it seems that her car is losing its grip on the road:

59 See D'Arms and Jacobson (2003).

> On a later occasion...with someone else driving at a slow speed on an isolated road, a very slight skid and the momentary sensation of uncontrolled movement had me gasping audibly for a second out of fear.... [I]t was the *thought* of danger or the tendency to call it to mind—something not quite explicit, but with clearcut behavioral (and physiological) effects.... This was based on a judgment that I made earlier, in the situation of my accident, presumably; but I deny that I actually extended that judgment, *as* a judgment, to the later situation, even momentarily. Instead, I would say that the judgment gave rise to a sudden thought, logically unconnected to my current beliefs. (1988:18-19, original italics)

Roberts, another leading proponent of the view, compares emotional experience to 'seeing' or 'construing' a situation in a certain way. Anger is the experience of 'construing' a situation as some offense that another has made against me or mine, guilt as one's having done something morally wrong, etc. A construal, as he puts it:

> is a mental event or state in which one thing is grasped in terms of something else...a perception, a thought, an image, a concept. I can perceive one face in terms of another, which I am also perceiving...I can perceive a face in terms of a concept, like *rugged* or *kindly*; I can imagine my living room in terms of furniture in the store, which I am presently perceiving; or in terms of either the image or the thought of my parents' living room, or in terms of the concept *grandiose* or *well-coordinated*.... (1988:190, original italics)

He talks at some length about the well known drawing that can appear as a duck or as a rabbit. Knowing that the duck-rabbit can be seen as a duck does not entail being able to see it as a duck. As with many images that are embedded in larger illustrations, one can know that they are there to be seen without being able at the moment to see them. Rather, seeing the duck-rabbit as a duck seems to be something like construing it as such, where this means something like "dwelling on or attending to, or at a minimum holding onto, some aspect, for example, the duckiness of the duck-rabbit" (1988:187).

What matters to the quasijudgmentalist thus is not what emotions are, but how we can talk about them. Whatever the emotions are (not judgments), we can identify them by their propositional contents. Emotions are 'thoughts' that a certain situation has obtained, or 'construing' a situation in a certain way. The view is right in line with the suggestion that emotions be individuated by their characteristic concerns; we can understand what they are just by filling out their propositional contents, and we can assess the emotions by evaluating whether they are felt in response to what they are characteristically concerned with. Greenspan, for example, describes her fear of sliding on an icy road while driving at a slow speed on an isolated road as "logically unconnected" from her current beliefs, suggesting that she knows it isn't appropriate to feel afraid in her situation. As well, we could 'construe' objects in ways to express properties that they don't really have, such as 'construing' the percept of a mouse on the kitchen floor and as something dangerous. The quasijudgmentalist removes herself from the debate over the ontological status of emotions by granting that they are not judgments, and retains only the method we use to identify them.

Emotional experience seems most aptly described in quasijudgmentalist terms. When feeling guilt, for example, it's very intuitive to say that we are 'construing' or 'regarding' a certain action-situation as morally wrong. Granted, this is a metaphorical way of describing things. But metaphorical talk of 'construal' or 'seeing as' is also commonplace in the philosophy of mind. Consider Dretske's distinction between two senses of looking or seeing, one that involves belief in some sense and one that doesn't. (He subscripts each sense of 'look' differently, the belief sense with a 'd' for 'doxastic' and the phenomenal sense with 'p'):

> To say that a dog $looks_d$ like a poodle to S is to say that, in the absence of countervailing considerations, this is what S would take the dog to be, what S's perception of the dog would (normally) prompt S to believe. Describing the dog as $looking_d$ like a poodle to S implies that S has the concept POODLE, understands what a poodle is, and classifies or identifies what she sees in this way...There is, though, another sense of these words, a sense in which if the dog looks the same to Susan as it looks to me, and it looks to me like a poodle, then it must look to Susan like a poodle whether or not she understands what a poodle is, whether or not she has the concept of a poodle. Following a long tradition, I will call this the *phenomenal* sense of 'look' ($look_p$). (1995:68, original italics)

Suppose that you're late with the rent and your landlord has been fussing about it. After arriving home from work one day you find a piece of paper nailed to your front door. What do you see? According to Dretske, there are really two different perceptions going on here. On the

one hand, you see a piece of paper nailed to your front door, which is in fact an eviction notice. In this sense, what you see is no different than what someone else with a relevantly similar visual apparatus would see. This is what Dretske would call the phenomenal sense of seeing or 'seeing$_p$.' On the other hand, you see an eviction notice as an eviction notice. Seeing in this sense involves having the concept EVICTION NOTICE. It also involves believing that you've been late with the rent, and that your landlord is motivated to take some action about it, etc. This is what Dretske calls the doxastic sense of seeing, or 'seeing$_d$.' Seeing an eviction notice as an eviction notice would normally prompt one to believe that there's an eviction notice nailed to one's door. More generally, seeing$_d$ involves bringing one's conceptual framework to bear on one's experience.

Now consider the emotion of fear. Suppose you're late with the rent and your landlord has been fussing about it. After arriving home from work one day you find a piece of paper nailed to your front door. You freeze. Terror seizes you. What's happening? It seems that your fear depends on your belief system in much the same way that seeing$_d$ the eviction notice depends on having the concept EVICTION NOTICE, as well as some other beliefs. If you did not know what an eviction notice was, or believe you had any reason to suspect that your landlord might serve you one, a piece of paper nailed to your front door would not elicit fear. Further, your fear would likely produce the belief that your impending situation is threatening.

Guilt too is amenable to a quasijudgmentalist interpretation. As a fitting response to moral transgression, guilt requires some conceptual hold on the difference between right and wrong. So when I reflect on the fact that I've taken a cookie from the cookie jar, I bring my understanding of WRONG to bear on what I've done; I construe my action as an instance of moral

wrongdoing. Without having any understanding of the difference between right and wrong I could not do this. Seeing$_d$ a morally wrong action is straightforwardly analogous to feeling guilt.

This all seems problematic for the sentimentalist. The nature of emotional experience is conceptual and involves evaluative concepts. As guilt is a distinctly moral emotion, it involves the concept WRONG and the threat of circularity and to the univocity of moral dispute looms. Nevertheless, despite the cognitive nature of emotion, guilt need not inherit the ambiguity of the concept WRONG. Indeed, we can assess emotions by a different criterion, one that does not involve evaluative concepts. Guilt is a feeling whereby one construes her situation as one in which she committed a moral infraction. The nature of this construing is the topic of the next section.

2.3 Guilt and Representation

Emotions have two components. On the one hand is the cognitive component. Emotions can be type-identified by their contents which are filled out by some propositional gloss. In this way, we speak of emotions having a characteristic concern. On the other hand is the affective component. Emotions have a feeling quality. Accordingly, distinctive emotion types have distinctive affects.[60] These too allow us to type-identify different emotion states. My aim in this section is to make sense of how these two components of emotion go together. Following a recent development (Prinz, 2004a), I argue that emotions are representations of certain value-laden features of the world. According to this view, feelings play a functional role in conveying

60 Affect I take to be different from an emotion's valence, which is the positive or negative quality of affect. Both guilt and shame, for instance, have a negative valence. That is to say, they both feel bad. But guilt feels different from shame—a difference which constitutes a difference in their affect.

information about values. Emotions thus carry out a cognitive task achieved in part by having a distinctive affective quality. This representational take on emotion allows allows us to type-identify emotions according to their affect in a way that's consistent with a quasijudgmentalist approach. Dispute over whether guilt is an appropriate response to Φ-ing can then be cast as a debate over whether a certain feeling-state is fitting. This allows differing parties in moral dispute to have different ideas of what counts as 'wrong' and yet have the same idea of what is meant by 'guilt.'

Individuating emotions by their feel corresponds to an idea that descends from William James (1884) and Carl Lange (1885). They argued that emotions are experiences or perceptions of certain physical states of the body. Fear, for example, would be the experience of being in a certain kind of physical state, something that might be accurately characterized by an increased heart rate, palpitations in the skin, raising of hair, widening of the eyes, shallowed breathing, etc. What characterizes fear is the experience of what it's like to be in this state. It's the feeling aspect of this state. James asks us to imagine an emotional experience without the feeling state: "if we fancy some strong emotion, and then try to abstract from our consciousness of it all the feelings of its characteristic bodily symptoms, we find we have nothing left behind, no 'mind-stuff' out of which the emotion can be constituted, and that a cold and neutral state of intellectual perception is all that remains" (1884:193). Indeed, emotional experience is essentially one of feeling.

Most notable support for this thesis is its corroboration by an extensive research program that has revealed several discreet physical states underwriting correspondingly distinctive emotional experiences. The program began in 1971 with the research of Paul Ekman and

Wallace Friesen. They conducted a series of experiments with the Fore, a preliterate people who lived in the highlands of New Guinea in complete isolation from western influence until 1959. The Fore were read stories designed to elicit certain emotional responses. One story described the death of a child, for example, while another tells of an encounter with an old friend. In addition to hearing these stories, the Fore were asked to choose a photograph of a facial expression to go with each story. One photograph shows that of a frowning face, another with a nose wrinkled, one with a smile, etc. Altogether there were six expressions to correspond to six different emotion types: joy, sadness, anger, surprise, disgust, and fear. By and large, the Fore gave answers that would be predicted of Europeans and Americans. They tended to associate expressions of surprise and joy, for example, with the story that described a chance encounter with an old friend, the expression of an upturned nose with the story of someone smelling something awful, and so on. Ekman and Friesen concluded that certain facial expressions were universal signs of emotional experience.

Ekman hypothesized that the regularities in observed facial expression among the Fore are underwritten by what he called 'affect programs.' Affect programs have three distinctive features. They are complex, coordinated and automated.[61] They are complex because they involve several different kinds of changes. Among them are: expressive facial changes, muskoskeletal responses such as flinching, endocrine system changes, and automatic nervous system (ANS) changes. Affect programs are coordinated in the sense that these changes occur in recognizable patterns. Finally, they are automated because they typically occur without requiring any conscious direction. Thus, the affect program is the coordinated set of changes, occurring without conscious control, that constitutes the emotional response. Facial expression

61 For a very helpful discussion on affect programs, see Griffiths (1997).

responses as those observed in the Fore suggest motivation-eliciting systems evolutionarily designed or socially constructed to represent states of affairs that have been of heritable significance in our ancestral past. As well, they often serve as cues, suggesting an adaptive role in coordinating behavior, and appear to operate independently of our ability to exert conscious control, being intimately related to experiential inputs. Further research into ANS, hormonal and muskoskeletal activity also supports this hypothesis.[62]

It's not enough to say that emotions are constituted by a specific set of bodily states, however, for it seems that emotions can occur without these corresponding states. Emotions can be disembodied, as it were. Consider first a nonemotional example, that of the phantom pain. Phantom pains are felt in the absence of any bodily concomitant. One can feel pain in one's leg without having a leg, for example. It's not accurate to say that this unfortunate person is not in pain, obviously, since she expresses all of the symptoms of pain, including self-reports. Rather, this person feels pain in a limb that does not exist or is not there. Related cases are those of emotional response that occur in the absence of bodily changes when brain centers ordinarily associated bodily change are stimulated. Just as sensory experiences can be produced endogenously, i.e., by stimulating areas of the brain that function as if the body had been perturbed, we can have emotional responses without actually experiencing relevant changes in the body.[63] But the James-Lange hypothesis anticipates that emotions could be disembodied. As the experience of a body state, rather than the state itself, emotions can occur without the corresponding physical correlate. Such experiences are rare but possible.

62 For surveys of this research see Griffiths (1997) and Prinz (2004).
63 Damasio (1994) called this the “as-if loop.”

Experiences associated with affect programs have a usefulness that descends from how that emotion has served us in our evolutionary past. Thus, fear has the function of delivering information about dangerous situations in virtue of its tendency to provide a selective advantage to those who felt it in situations that proved legitimately threatening. The feeling is associated with behavioral tendencies to freeze or flee from such situations, and those who reliably responded to them with fear tended to survive and pass this tendency along to their offspring.

There is a rationale for feeling guilt that bears some similarity to how animals interact. Animals do appear to abide by norms that coordinate their interactions with one another. Two dogs meeting on neutral ground will engage in certain rituals and behaviors. These appear to be governed by norms that help determine what's to happen in open-ended territorial disputes. As Gibbard remarks, "animal interactions follow certain regular patterns, and the patterns seem, in a way, to have a rationale...they constitute adaptions, we may presume; that is to say, they are the result of natural selection favoring these patterns" (1990:69).

Consider also Frank's speculative account of how guilt could have evolved to solve 'commitment problems.' A commitment problem arises when it is in a person's interest to behave in ways that will later be contrary to one's best interest (1988:47). For example, suppose two people could enter into a profitable venture together. Say that each will have the opportunity to cheat without the other noticing. Regardless of what the other does, it's in the best interest of each to cheat. However, if both cheat, both do worse than if neither cheats. If each commits now to open and honest behavior they will later miss out on opportunities to cheat without any significant chance of getting caught. So how does one decide to act in ways that one knows will be contrary to one's best interest? One way to solve this problem is by being disposed to keep

one's promises. Guilt could serve in such a disposition. If each were disposed to feel guilty for going against one's word, then the other could trust that they will not do so. Of course, the kind of coordination taking place on this level of explanation is different than that between dogs. But the basic point is just the same. Guilt may be an adaptive feature of human beings that derives from the basic need to coordinate social behavior.

However, its distinctive social application suggests that guilt does not appear to be something that we have received from non-human animals. Many are convinced, for instance, that the fact that guilt requires some conceptual hold on right and wrong implies that it does not exist in animals (Hauser 2002) or infants (Lewis 2000). And most developmentalists speculate that guilt emerges no earlier than the second year of life. One might find it telling that dogs often appear to express guilt. Scolding Fido for his poor indoor behavior may elicit a guilty look, perhaps a tail between the legs, droopy eyes, etc. But this behavior may be learned. Further, revealing evidence for a genetic predisposition for guilt would have to come from primates. And, so far as I know, there is no conclusive evidence that primates feel guilt.[64]

Guilt is what Griffiths (1997) calls a 'higher cognitive emotion.' Unlike affect programs, guilt represents moral transgression. This serves as a basis of explanation for a range of differences between affect programs and guilt. Consider:

> The stimulus appraisal which initiates an affect program reaction is to a large extent informationally encapsulated. The subsequent complex set of actions unfolds automatically, and it is difficult to interfere with these actions voluntarily. There are a large number of emotions which do not conform to this model. In many instances of

64 On this point, see Prinz (2004a:127)

> guilt, envy, or jealousy the subject does not display a stereotypical pattern of physiological effects. In addition, these emotions seem more integrated with cognitive activity leading to planned, long-term actions than the affect program responses." (Griffiths 1997:100)

Affect programs are 'informationally encapsulated;' they carry information about one's environment that cannot be impeded by conscious control. Fear can be elicited by certain visual images, such as that of a coiled snake. Encountering this percept will elicit fear on most occasions, regardless of what one believes about the danger of these creatures. Yet there is no corresponding universally recognized elicitor for guilt. Furthermore, even if there were guilt-eliciting percepts, it seems that we could change our responses to them by changing our beliefs about what counts as 'wrong.' Say, for example, that I formerly believed that stealing from the cookie jar is wrong, while now I do not believe that it is wrong. It's not difficult to imagine that I can now steal a cookie from the cookie jar without feeling the slightest bit guilty. I don't think it's wrong to do after all. The image of a coiled snake, however, will cause a certain physiological reaction on most occasions, regardless of how dangerous I think snakes really are.

Guilt also appears to be related to long-term planning in ways that affect programs are not. If, for example, after stealing a cookie from the cookie jar I feel guilty, I may plan to make amends for my bad deed. Perhaps I will ask forgiveness from my mother. Maybe I will confess my sin to the local pastor, strive to become a better person, and do my best to ensure that others don't tread the same unfortunate path that I chose. An affect program, on the other hand, has implications only for my immediate behavior. If I feel this variety of fear, I might flee from the

threatening object, or fight. But that is all that appears directly related to the fear I experience. As an affect program, my feeling is occurrent and fleeting. It likely will bear no significance in how I conduct my future plans (except perhaps to keep a careful lookout for snakes).

Guilt requires having some idea about right and wrong. But whatever this means, it should fall short of judgmentalism. Guilt does not require the judgment that one has done something wrong. Thus, guilt is not a cognitive sharpening of some other basic emotion. The kind of recalcitrance that occurs with cognitive sharpenings is much different than what happens with guilt.[65] Take tenure rage. A recalcitrant version of tenure rage would be feeling rage at the senior faculty without believing that one has been denied tenure. Notice how similar this is to holding contradictory beliefs. The victim is here both angry that he has been denied tenure while believing that he has not been denied tenure, what appears to be an extremely irrational psychological state. Recalcitrant guilt simply involves feeling guilty for something one does not think wrong to do. The latter is much less like holding contradictory beliefs than the former.

Following Prinz (2004a), I propose that guilt is a specific kind of feeling-state that has the function of detecting those actions that correspond to our beliefs about what is wrong to do.[66] As such, it has two salient features (Dretske 1995). It carries information about some object and has the power to misrepresent that object. A flagpole, for example, carries information about the temperature of the air. But a flagpole does not represent the temperature of the air, since it cannot misrepresent the temperature of the air. A thermometer, on the other hand, can represent the temperature of the air because it both carries information about the air and has the power to misrepresent the temperature of the air. The thermometer can indicate that it's '32º' when it is not

65 On this point, see also D'Arms and Jacobson (2003:142).
66 Thus the current proposal is a version of quasi judgmentalism.

32º. The flagpole cannot inform us incorrectly of the temperature of the air. Representational states thus have the function of indicating the properties of other states. Flagpoles do not have the function of doing so. Thermometers do. Thermometers (outdoor ones) have the power to misrepresent the temperature of the air because they have the function, i.e., are supposed to carry correct information about it.

Guilt is like the state of mercury in a thermometer. If functioning correctly, it indicates the presence of what we believe to be a moral infraction. Accordingly, it will misrepresent its object under two conditions. First, guilt misrepresents its object if that object does not in fact correspond to what we believe is wrong to do. Second, guilt represents moral transgression in virtue of the function that we have given it to do this. It is we, through the establishment of certain norms of behavior, who have set the attitude's purpose.[67] Further, this model is quasijudgmentalist in the following way. We can say, very loosely, that the feeling of guilt is the body's way of construing an action as morally wrong (though guilt can be disembodied). It's a way of responding to a situation in a way that brings it under the concept WRONG. This doesn't mean that one who feels guilty actually believes that he has done anything wrong. He need not believe this any more than a thermometer 'believes' that it's 32° outside. The thermometer doesn't actually judge anything to be the case but simply has the function of carrying certain information about the world.

This schematic way of putting the matter leaves it open exactly how guilt has the function that it does. It's an open question, for instance, whether guilt is a blend of two more basic emotions, whether it's a specially derived modification of some more basic emotion (Prinz 2004a), or whether it should be examined in some other way. Griffiths (1997) thinks that all

67 These remarks have important implications for the fittingness of guilt. See Chapter Three.

higher cognitive emotions comprise their own category and do not form a natural kind with responses like affect programs. But whatever the most plausible story turns out to be, it should correspond to two restrictions: guilt is not tied necessarily to moral judgment and is a distinctive kind of feeling-state.

A representational theory of guilt might be developed in a way that effectively denies the latter claim. Prinz, for instance, believes that guilt is a recalibrated form of sadness (2004a). States are recalibrated when they are put to a different use than what was originally intended. Consider the thermometer example again. Suppose that the state of a thermometer registers '32º.' The state that it's in, call it S, carries information about the temperature of the air in virtue of its role in a system. Mercury contracts in cooler temperatures and expands in warmer ones and we can measure changes in the temperature by correlative changes in its density. Further, we could take this measure incorrectly. If the system was damaged in some way, for instance, if the glass was broken perhaps, we might get an incorrect reading. Now suppose that, instead of '32º,' we mark S as 'freezing.' Now, the thermometer has the function of indicating the temperature of the air when it is freezing. The reason it has this function is because we give it this function. S now indicates when it is freezing outside not in virtue of the system in which it is a part, but rather in virtue of the type of which it is a token (Dretske 1995). If we give S the label 'freezing,' then every time the thermometer experiences S, it purports to report the presence of a particular value, 'freezing.'

The same state can used for different purposes and so represent different values. On one thermometer S is labeled '32º,' on another S is labeled 'freezing.' Both thermometers have the same state but represent different values. The difference is a difference in tasks they have to

perform. It makes sense, then, to talk roughly of 'beliefs' that these thermometers have. On the one hand, one thermometer 'believes,' i.e. represents, that it's 'freezing' outside. It has been given the task of indicating when the temperature drops below 32º Fahrenheit, a range considered 'freezing.' Another thermometer, however, 'believes' that, in S, it is '32º.' The information it has the function of conveying is different. Both, however, 'experience,' we might say, S. Both have the same 'experiences' but different 'beliefs.'[68]

Emotions like anger, sadness and fear, have a phylogenetic history which tells of their role in conferring a selective advantage to our species. Fear has been designed to detect danger by evolution. Predators and threatening situations have played a role in the adaptive success of our species. The better we were able to avoid them or defeat them, the more likely it was we passed on our genes to our offspring. The upshot has been a physical response to dangerous things that prepares us for avoiding or taking on the threat.[69] The state of the body corresponding to fear, call it F, carries specific information about the world. F has the function of telling us that the situation we are in is dangerous. F has this function in virtue of the role that evolution has given it to play in the human body.

Guilt would be different. As a recalibrated form of sadness, guilt would have the function that it does not in virtue of being part of a system but because we have simply given it this function. On this account, guilt is the experience of sadness recalibrated to pick out moral transgression. Where our ancestors used sadness to relay the presence of loss, we use sadness at times as a sensitivity to morally wrong actions. This new use for sadness has been achieved

68 This is not to say of course that thermometers actually believe anything. They are not minds in the possession of concepts. This is why I present the term in scare quotes.

69 The fight or flight response.

through our moral education and reinforced by standards that call for feeling it on occasions that correspond to what we believe is wrong to do.

However, the idea that the same feeling-state could occupy different roles is implausible. Most would agree, I think, that sadness does not feel exactly like guilt—a fact that confounds the suggestion that both emotions are underwritten by the same feeling state. Further, supposing this distinction makes ready sense of the telling discrepancy between the behavioral tendencies of guilt and sadness. Guilt elicits the desire to apologize, to agonize about one's past behavior and to avoid future alternatives for wrongdoing. Sadness propels one to mourn. The reason for these specific behavioral tendencies seems to derive from the fact that our feelings themselves propel us to these different ends.

On the recalibration model, the distinction between these behavioral tendencies would be cashed out in terms of the concerns embodied. That is to say, it would be the propositional content of these emotions that explains the sensitivity the one feeling-state has to different kinds of situations. But this would reintroduce judgmentalism. Indeed, there would be nothing 'quasi' about this picture. Feeling sadness could not be just to experience a situation as an instance of loss, since that is identical to feeling guilt. Rather, sadness would have to be the feeling of guilt plus the judgment that one has incurred a loss. Instead of construing emotions as a single state with two aspects, the recalibration model analyzes them into two separable components.

On the representational thesis offered here, guilt has two sides. It has a propositional content that can be filled out with the moral concept WRONG. Guilt has the function of detecting whatever set of natural properties fall under this concept. In this way, the feeling is a construal of one's environment in a moral light. Again, the experience is not equivalent to making a

judgment *per se*. Rather, the sense in which guilt is a judgment is just that in which it has the function of detecting whatever properties constitute what one believes is a moral infraction. The other side to this story is the distinctive feeling-aspect of guilt. An essential part of guilt's function is achieved by how it feels. Indeed, the presence of moral wrongness is conveyed to the agent in virtue of the way guilt feels.

What we gain from this take on guilt is a point of access to type identify it independently of our concept WRONG. This disarms the threat of circularity and provides a way of accommodating the essential contestability of moral concepts alongside their univocity. Parties with disparate concepts of WRONG can be moored in a discussion of whether guilt is fitting because guilt has a distinctive feel and can be referenced by that feel. Thus, when two people disagree about whether Φ-ing is wrong to do, we can interpret their disagreement as one over whether it is fitting to feel 'this way,' whatever it is like to feel guilty.

Further, the propositional side of guilt explains the extent to which people with different concepts of WRONG are lodged in an intractable dispute. When they do, the dispute takes on the impression of a kind of pointless quarrel. It's as if each gets going in their own circle, defining 'wrong' in terms of 'guilt' and 'guilt' in terms of 'wrong.' Nevertheless, the argument is a real one. They are not simply talking past one another. To capture this aspect of this situation we need the other side of guilt, its feeling-aspect.

Judgments about guilt's fittingness bear in important and significant ways on how we behave. Accordingly, the representational theory of guilt outlined here plays a crucial role in the explanation I ultimately offer. Guilt's representative powers pave the way for understanding how it functions in what we believe. Its connection to our conceptual hold of wrongdoing links it to

our judgments while its feeling-aspect connects it to our motivations. Thus, the emotion serves as a kind of mediator between judgment and motivation. And as it turns out, this is exactly what we need in order to explain the relationship between what we believe we ought to do and what we are motivated to do.

Chapter Three

Moral Motivation

3.1 The Problem

> Mabel and Fred want to marry each other. The opportunity is there, the desires are aflame, the consequences are predictably acceptable or even desirable. There is only one thought to oppose it: they have a duty to do otherwise, so it would be wrong. And this feeling that it would be wrong can wrestle with and sometimes even overcome all the rest. Isn't this mysterious?
>
> --Simon Blackburn (1973:154)

Believing that something would be wrong to do typically entails being motivated not to do it. Those who try to explain this phenomenon generally fall into one of two camps. In one camp stand internalists, who believe that the connection between moral belief and motivation is analytic. They hold that one who is not motivated to do what they believes is right simply does not understand what they are saying, does not really say what they think they say, or is somehow irrational. Externalists, on the other hand, think that the connection is due to contingent psychological facts about us, our moral education for instance, the norms and values that society

instills in us to ensure a well functioning social order. Contrasting intuitions give rise to disagreement over whether it's possible to believe doing something would be wrong while at the same time not caring at all—internalists say it's not while externalists disagree. Moral motivation thus has often been construed as the debatable thesis that moral belief and motivation always coincide.[70]

Focus on this debate, however, can make it difficult to parse out the phenomenon both are trying to explain. What separates the two is not so much the fact that they disagree on this score but rather why they disagree. While the internalist insists that moral belief and motivation always coincide, the externalist is content to hold that one could believe that something was the wrong thing to do without feeling the slightest impulse to resist doing it, if certain psychological facts about us had been otherwise. Given, however, that we are who we are, there is a connection between what we believe and what we're motivated to do. Both agree that motivation and moral belief have some interesting connection. What they disagree about is the nature of this connection.

A similar way in which the debate has been clouded, I think, is due to controversy over the status of the so-called 'amoralist.' The amoralist is one who believes that something is wrong to do but cares not at all about that. Say, for example, an investment banker descries an opportunity to embezzle some money from the firm that he works for. Taking sufficient time to work through all of the details, he concludes that he can accomplish this without risking any significant danger of being caught. Of course, he thinks, such a thing would be wrong. It's not his money after all, and he doesn't need it. But it would allow him to take exotic vacations during his summer weeks off, and fill his home with objets d'art, etc. These things would be

70 See Brink, (1989), Smith (1994), Mele (1996) and Dreier (2000), to name a few.

good for him at least, and that's all that's important to him. It would be wrong, he confesses, but so what?

Neither externalist nor internalist denies that this person is in some way unusual. We would expect the investment banker to be at least bothered by the fact that he would be doing something wrong. A person who has no inclination whatsoever to do the right thing does not seem properly responsive to moral considerations. Some have even thought that the amoral person is psychologically impossible. All of this only underscores the point that moral judgment typically indicates having at least some motivation to act accordingly. Both internalist and externalist agree on this much. The fundamental disagreement between them, again, is why. It's their different intuitions about the nature of moral motivation that brings them into disagreement about whether moral belief and motivation must always go together.

The coincidence of motivation to moral belief is most clearly expressed by the fact that changes in moral belief are reliably tracked by changes in one's motivations. If one changes her beliefs about what it is wrong to do, then we should expect her motivations to change along with them. We would observe this kind of tracking only if motivation and moral belief typically occur together. Here's an example given by Smith:

> Suppose I am engaged in an argument with you about...whether we should vote for the libertarian party at some election as opposed to the social democrats.... Suppose that I come to the argument already judging that we should vote for the libertarians, and already motivated to do so as well. During the course of the argument, let's suppose you convince me that I am fundamentally wrong.... You get me to change my most

> fundamental values. In this sort of situation, what happens to my motives?...If I am a good a strong-willed person then a new motivation will follow in the wake of my new judgment. (1994:70-71)

Smith's observation is about 'good and strong-willed' people, a phrase that in later writings he expresses regret for (1997:111, n.27), preferring instead the term 'moralist,' which is a better term to use. Imagine someone who believes at one time, correctly, that discrimination against the Jews is morally wrong. Suppose also that he is motivated to speak out against those who advocate treating Jews in ways that would suggest they don't have the same rights and privileges as others, and so forth. But now suppose that this person becomes convinced, through long and arduous argument, that his belief is actually incorrect. Now, he believes that it is morally acceptable to discriminate against the Jews. In fact, he might even be convinced that it is his moral obligation to treat the Jews as people less deserving of the rights and privileges others enjoy. He is motivated to speak out in favor of those who think the same, and against those who believe that discrimination against the Jews is morally wrong, etc. But his new belief is incorrect. This person, it seems, is not a 'good' person. He believes what he does is good, because he thinks that his moral beliefs are true. But they aren't. At least, his beliefs about how to treat the Jews aren't. This person is not, then, 'good and strong-willed' but rather takes himself to be 'good and strong-willed.' He is a moralist.

The motivation-tracking phenomenon occurs in moralists, which is to say, in most of us. Most of us believe ourselves to be correct in our moral views. Most of us would also find it somehow inconsistent if our beliefs were to change and our motivations didn't. Most of us

believe, simply, that we ought to do the right thing. And if we come to believe differently about some moral issue, we will probably change the way we behave. In as much as we are 'good and strong willed people,' so we say to ourselves, we would be motivated to do what we later and correctly believe is the right thing to do.

Motivations will track changes in our beliefs about what it is morally right to do. Where I formerly believed that voting for the social democrats is right to do, I now believe that voting libertarian is right to do. More precisely, my beliefs are about what it would be morally wrong for me not to do. I formerly believed that it would be morally wrong of me not to vote social democrat; now I think that it would be morally wrong of me not to vote libertarian. The sense in which it is 'right' for me to vote libertarian is the same sense in which I morally 'ought' to Φ. Notice too how we would also expect motivation to track moral belief when we change our views about what is morally permissible to do. Say I formerly believed that abortion was morally wrong to do, but after listening to several arguments *pro* choice I now believe that it is morally permissible. My disposition to speak out against it, to resist having one, etc., it seems would change. So our motivations should track changes in our moral beliefs in both directions. Where our beliefs about what is right to do change—either in terms of 'morally right' or 'morally permissible'—so too should our motivations.

However, though our motivations usually track changes in our moral beliefs, they don't always. I may be convinced that voting libertarian is what I ought to do but still be motivated in the end to vote social democrat. Dissonance between belief and motivation can occur in various ways. It can be persistent, as in cases of listlessness—where there appears to be something wrong with the agent who experiences a complete lack of feeling for what he thinks he ought to

do. It can be gradual. We should expect the tracking phenomenon to occur gradually over time. If I formerly believed that abortion was morally wrong but come to change my mind, I will likely continue to experience fits of resentment for those who undergo the procedure, perhaps even the temptation to criticize *pro* choice openly, even though I myself am now *pro* choice. The gradual nature of change in the motivation tracking phenomenon seems to be something essential to moral agency. As Blackburn remarks, "we cannot become corrupt overnight, and usually we cannot tell when we have done so. Indeed, it would be a hallmark of many kinds of moral blindness that this is so. The really coarse man thinks that he is perfectly in order, but that other people are too fastidious (recognizing that you have become really coarse is in this way self-refuting: the realization itself shows some residual delicacy)" (1973:160).

Another distinct possibility is an isolated fit of listlessness in otherwise normally disposed persons. The stress of one's job might prevent one from being motivated at all to care after one's obligations. Indeed, listlessness can occur for reasons we are all familiar with. As Stocker famously puts it,

> Lack of this desire [to do what one believes is right] is commonplace. Through spiritual or physical tiredness, through accidie, through weakness of body, through illness, through general apathy, through despair, through inability to concentrate, through a feeling of uselessness or futility, and so on, one may feel less and less motivated to seek what is good. (1979:744)

It could happen that any of these causes arise in those normally disposed to do what they think they ought to do. Moreover, typical syndromes of listlessness may prevent one from having any motivation at all to act in accordance with their convictions.

This suggests that the motivation-tracking phenomenon occurs in those who are not suffering from listlessness or some other sort of psychological malady. The problem of moral motivation thus involves the effort to explain why this would be true. Let's turn now to a more thorough investigation of the problem and how it might be solved.

3.2 Internalism and Motivation

In this section I present the general internalist thesis about motivation. I examine several incarnations of the view, refining it to accommodate various intuitions of belief-motivation dissonance. In the end, I arrive at a thesis that I ultimately want to argue for. The discussion in this section is an effort to bring out the problem of moral motivation and indicate some direction on how to solve it.[71]

Consider first Nagel's description of motivational internalism, which serves as the starting point for many discussions of this issue. He writes:

71 The idea that moral belief and motivation are modally distinct may seem conceptually foreign to noncognitivists about moral judgment. Mele (1996), accordingly, thinks that the scope of the debate should be limited to moral cognitivism. But this is too severe. Even noncognitivists respect what appears to be a separable relationship between moral belief and motivation. Gibbard (1990), for instance, analyzes moral judgment in terms of an expression of one's acceptance of a moral norm. But he allows that one can 'accept' a norm without 'internalizing' it—where this means having one's motivations in concert with what one 'accepts.' The remarks I make here are meant to be very general, applying to cognitivists and noncognitivists alike.

> Internalism is the view that the presence of a motivation for acting morally is guaranteed by the truth of ethical propositions themselves. On this view the motivation must be so tied to the truth, or meaning, of ethical statements that when in a particular case someone is (or perhaps merely believes that he is) morally required to do something, it follows that he has a motivation for doing it. (1970:7)

At least two distinct theses are discernible here. First is the claim that if someone is morally required to Φ, then she has a motivation to Φ. Second, it is suggested that that if someone believes that Φ-ing is the morally right thing to do, then she is motivated to Φ. The first thesis is a motivations internalism about moral obligations. The second is a motivations internalism about moral judgments.

The first thesis is false. One could very well be morally required to Φ and yet have no motivation to Φ. This could happen if one's moral beliefs were incorrect. Consider someone who believes that black cats are bad luck—that they infect with ill fortune anyone who associates with them. Suppose that this person believes, further, that it is the moral duty of anyone who sees or knows of the presence of a black cat to destroy it. But it is not morally right to do any such thing (suppose). This person's moral beliefs are based on a false belief, namely, that black cats bestow bad luck on people they come into contact with. But because of this false belief, and the corresponding false moral belief, this person has no motivation to do the right thing. Thus, despite the black cat hater's moral obligation not to kill cats, he does not have the motivation to comply with it.

The second thesis about judgments seems true. Notice that it would be odd if the black cat hater wasn't motivated to destroy any dark feline he found out about. And this seems to be because he believes that he has a moral obligation to destroy black cats. So the more plausible thesis seems to be the one about judgments. Our motivations to do the right thing don't seem to derive from moral obligations themselves, but from our beliefs about what those obligations are. The internalism I'll be assessing is one about beliefs. As a first rendering, consider I:

I: Belief that Φ-ing is morally right entails being motivated to Φ.

An action can be morally right in any of three senses. First, an action can be right in so far as being an alternative to a morally wrong action. In this case, moral right means something like 'morally obligatory.' Consider a case of wanton cruelty. It would be both wrong to engage in such an action, and morally right not to. We have a moral obligation not to commit actions of wanton cruelty. Second, an action can be morally right if it is morally permissible to do it. In this kind of case there is nothing one ought to do, morally speaking, but one or several things one could do without violating a moral obligation. And of course there are countless examples of such actions, such as going to the grocery store, watching a movie, or whatever. Third, an action can be morally right in the sense of being supererogatory. Supererogatory actions go above and beyond the call of duty. Think of giving to charity when one has little money. Giving to charity in such circumstances is morally right to do, but not necessarily morally wrong not to do; Rather, it achieves something more than what is morally required.

I is most plausibly a claim about believing that Φ-ing is morally right in the first sense. This is evidenced by the all too familiar conflict between what we want to do and what we think we morally ought to do. Morality operates as a constraint upon alternatives for actions and we typically resist the ones that morality forbids. Consider, on the other hand, beliefs about actions that are morally permissible. There's no special connection of motivation to beliefs about these kinds of actions. I may believe that going to the grocery store is not wrong of me to do, but this doesn't say anything about my motivations. I could go to the grocery store or to the post office, or not go anywhere. Each is morally permissible but none have any motivational import. Supererogatory actions, likewise, don't entail motivation. It's not difficult to imagine someone with little money having no motivation to give to charity even though she thinks it would be right. The relevant kind of moral belief is that of believing wrong to do or right not to do. It's these kinds of beliefs that are motivating, not those that are morally right in the sense of being supererogatory or morally permissible. I should be revised, this time in terms of believing morally wrong:

I1: Belief that Φ-ing is morally wrong entails being motivated not to Φ.

I1 is appealing in two ways. First, it has theoretical appeal. Asserting an internalism about motivation supplies the easiest and simplest answer to the question of why motivation and moral judgment go together. We observe this correspondence, according to the internalist, because moral judgment motivates as a matter of conceptual fiat. Second, it seems plausible. If someone were to say that he believes some action or practice morally wrong, but is indifferent to

the idea of doing it, we would tend to think that he doesn't really believe that said action or practice is morally wrong. Recall an earlier example. Say a temp-agency manager asserts the belief that sex discrimination in the workplace is morally wrong, but regularly harasses his female employees. It's likely he's just paying lip service to a value everyone generally accepts. He knows that he must appear to accept it lest he be regarded as sexist. This guy doesn't really believe that women are due totally equal treatment in the workplace, he just knows to say he believes this so that his character remains in good public standing. He is insincere in his belief.

But is I1 true? It seems not. Sociopaths appear to be everyday counterexamples. Sociopaths seem not to lack any of the essential cognitive assets necessary to score well on I.Q. tests. They appear to be able to think rationally and without delusion. They seem to understand social norms and mores well enough to anticipate how others will react to what they do. They comprehend the consequences of their actions, for themselves and others. Yet they seem not to care at all about what those consequences are. They can do terrible things to animals, other people, even friends and family, while expressing not the slightest bit of remorse or sympathy.[72]

However, there appears to be sufficient evidence suggesting that sociopathy effectively impairs the belief-state. Either sociopaths believe only in an inverted commas sense, or they are irrational in some sense. Consider the first alternative, that they believe only what other people believe to be wrong. In many cases, the sociopathic condition seems to have little to do with an affective disorder. Rather, it appears that they are unable to assign proper weights to the interests of others and even themselves. That is to say, sociopaths do not appear to understand how another person's interest or their own future interests could be a reason for them (the sociopath) to act. Kennet reports that the sociopath "fails to form any coherent conception of his own or

72 For a seminal study of the sociopath, see Cleckley (1964).

others' ends, and therefore the ways those reasons generate and sustain reasons over time, reason has only a tenuous grip on him. His actions are intentional, and perhaps short-term instrumentally rational, but that is all. Unsurprisingly, then, he is not troubled by cognitive dissonance when he makes inconsistent judgments about what he may do and about what others may do, since he does not care about, or does not understand, the point of rational justification in the first place" (2002:16).

If Kennet is right, sociopaths do not grasp the concept WRONG. They do not seem able to appreciate any concern except what is in their own current interest. They act on impulse, and are simply not bothered by the thought that what they do may have harmful effects on themselves or others. If this is true, then the best we can say about their belief that Φ-ing would be wrong to do is that they believe in an inverted commas sense—that when they use the concept WRONG they merely state what other people believe to be WRONG.

Now consider the second alternative: even if there are examples of sociopathy that do not entail inverted commas sense belief, there's some sort of irrationality at play. Recall the example of the investment banker who embezzles money from his firm. This person appears to appreciate how his reasons for stealing pertain to his long-term best interest. Indeed, there are probably many folks just like the embezzler—people who are shrewd and calculating but who have no concern for anyone's welfare except their own. For the shrewd sociopath, it's these considerations, ones that bear on how their actions affect others, that don't get involved in deliberation about what to do. The investment banker understands that what he does is wrong but doesn't see that as a reason not to do it. He doesn't care about what he does not because he

doesn't know what it means for something to be a reason to Φ. He doesn't care because he doesn't think that moral reasons are legitimate reasons to Φ.

The corrupt investment banker is missing something. Moral reasons are perfectly good reasons. They don't count towards what is in one's best interest necessarily. Φ-ing might benefit someone else at great cost to oneself. Still, moral action can be rational for all that. It's just that what counts for Φ-ing here are the concerns of someone other than oneself. Granted, it may be irrational to do what's morally obligatory, if the cost to oneself of Φ-ing is too great. But Φ-ing could be rational just because it's morally right.[73]

Thus, sociopathy is not a counterexample to I1. The condition involves either inverted commas belief or a kind of irrationality that fails to appreciate the normative strength of moral reasons. In the former case, moral belief does not have the right content. In the latter case, moral belief doesn't carry its normative weight. Either way, it seems that we can say that sociopathy entails a corrupted belief-state, enough to conclude that sociopaths do not genuinely believe (not necessarily insincerely) that Φ-ing would be wrong to do.

Genuine moral belief must not be insincere, an instance of inverted commas sense belief, nor accompanying a numbness to the normative force of moral reasons to act. One who believes in any of these ways may not be motivated appropriately. One won't be so motivated because one doesn't really believe that Φ-ing would be wrong to do. The sociopath does not have the practically rational equipment to genuinely believe, while the sexist who avows gender equality is only repeating what he thinks other people want to hear. These are not cases of genuine belief. For cases of genuine belief, according to I1, motivation is necessary.

73 I argue briefly for a pluralistic view of reasons in section *3.3.1*

I1 also has predictive value. It says that I will be motivated to resist doing what I think is morally wrong. So it would predict my motivations change when my beliefs about what is morally wrong changes. I1 also predicts, albeit with less certainty, that motivation would track changes in one's belief about what it is morally permissible to do. Say that I formerly believed that one should not vote libertarian, but now, after hearing several arguments in favor of the position I understand how someone might be sympathetic towards the position. I now think it morally permissible to vote libertarian. According to I1, I will not be motivated to resist Φ-ing in so far as I don't believe that Φ-ing is wrong. Since I don't think that voting libertarian is morally wrong, I will probably not be motivated to resist voting libertarian.

It could happen that one be motivated not to Φ despite believing that Φ-ing is morally permissible. Say, for instance, that I don't believe that I ought to read *War and Peace* (I believe it's morally permissible not to). However, given my disdain for Tolstoy, I am motivated not to read it anyway.[74] Thus, I1 leaves it open that one be motivated not to Φ while believing that it is morally permissible to Φ.

I1 also allows for changes in one's motivations to be gradual. Motivation can be read strongly or weakly (Brink 1989). If we read motivation in a strong sense, my motivation moves me to vote libertarian where formerly it moved me to vote social democrat. If we read motivation in a weak sense, then I am motivated to vote libertarian but may not vote libertarian because of some other motivation (say I still really like the social democrats). On a strong reading, motivation is overriding, while on a weak reading motivation is not necessarily overriding. I1 can also be given a strong and weak reading. On a strong reading, belief that Φ-ing is morally right entails having an overriding motivation to Φ. On a weak reading, belief that

74 I like Tolstoy actually.

Φ-ing is morally right entails having some motivation to Φ. Let's examine these two versions of I1 separately. On a strong version of I1, we get:

$I1_s$: Belief that Φ-ing is wrong entails having an overriding motivation to resist Φ-ing.

But $I1_s$ is false. Consider a variation of the case of the sexist temp-agency manager. Imagine that the temp-agency manager, call him Tim, is actually a good guy, but he's complex. Tim believes that gender discrimination is wrong, but was raised in a family and culture which relegated women to a subservient role. One of Tim's employees, Sue, is a good employee. Sue is due for promotion, and she's in competition with another fellow, Jim, who's not quite as competent as Sue. Given the influence of Tim's moral education, Tim wants to hire Jim, even though he now believes that doing so would be unfair to Sue. The thought that he would be doing something unfair to Sue bothers Tim, but in the end, Tim acts on the desire to promote a man over a woman and hires Jim. Tim feels terrible about this afterwards. Tim believes that sexism is wrong but is motivated, in the end, to act in a sexist manner. This isn't to say that his belief about the moral status of his action does not affect him. It does. Tim has some motivation to do what he believes is right even though this motivation is outweighed by a competing motivation to do otherwise.

$I1_s$ would also not allow for our motivations to gradually track changes in our moral beliefs. If, for example, I become convinced that voting libertarian is what I morally ought to do, then, by $I1_s$, I should be motivated in the end to cast my vote accordingly. My motivation all things considered should change as soon as I take on this new belief. But this isn't likely to

happen. More than likely, I will retain much of my concern for the cause of the social democrats, and have to fight against what I am accustomed to doing in order to do what I believe I ought. $I1_s$ anticipates that our motivations will track changes in our moral beliefs, but too well.

$I1_s$ thus overstates the connection between moral belief an motivation. Motivation may be a necessary condition for believing morally right. It is not, however, necessary that one be motivated in the end to comply with what one thinks one ought to do. Rather, if an internalism about motivation is correct, it would seem only necessary that one have some motivation to do what one believes is morally right. Consider now $I1_w$:

$I1_w$: Belief that Φ-ing is morally wrong entails having some motivation not to Φ.

Unlike $I1_s$, $I1_w$ allows for changes in our motivations to gradually track changes in our moral beliefs. According to $I1_w$, when I change my mind about who to vote for, I only have some motivation, necessarily, to vote for whom I now think I ought. Thus, it could happen that when I change my views about what to do I commonly find my motivations at odds with one another. While my former concerns persist, my new concerns are not overriding. It would likely take time for my newly acquired motivations to outweigh my old ones.

However, $I1_w$ is also false. There are cases in which one sincerely and without confusion believes that Φ-ing would be wrong to do and is not motivated at all to resist Φ-ing. Mele offers just such a case:

> Consider an unfortunate person—someone who is neither amoral nor wicked—suffering from clinical depression owing to the recent tragic deaths of her husband and children in

> a plane crash. Seemingly, we can imagine that she retains certain of her beliefs about what she is morally required to do...while being utterly devoid of motivation to act accordingly. She has aided her ailing uncle for years, believing herself to be morally required to do so. Perhaps she continues to believe this, but now is utterly unmotivated to assist him. (1996:733)

Listlessness can occur if one were very depressed, or suffered from some other psychological malady. These kinds of cases are different from those in which one does not really believe that something would be wrong to do. The best we can say about sociopaths is that they believe in an inverted commas sense or don't fully understand the reasons that are generated from their conviction that Φ-ing is wrong. People who believe insincerely are not motivated either because they do not actually believe what they say they do. Neither are counterexamples to $I1_w$. But it is possible to have genuine moral belief without motivation. In these cases, we can say that they know what it's like to be motivated to do what's right, but, for whatever reason, just aren't. These people are counterexamples to $I1_w$.

The internalist might insist that there is a conceptual connection between moral belief and motivation holds in normal people under normal circumstances (Dreier 1990). She might admit that it is possible that one genuinely believe that something would be morally wrong to do and yet not be motivated at all to do it, but insist that these cases simply reveal the contrast between what is normally true of individuals and abnormal instances of moral belief. This contrast provides us with some notion of normality in agents who make moral claims, and in such instances there is a conceptual connection between belief and motivation.

But what is normal? The phrase 'normal people under normal circumstances' cannot simply describe those who are motivated by what they believe is the right thing to do, for this would just define 'normal' in terms the internalist insists upon. What the internalist needs is an independent concept of NORMAL. One way to do this would be to isolate a distinctive feature of moral judgment in those whose motivations fail to correspond. If it turned out that under these conditions people were insincere or believed in an inverted commas sense, we would be able to say that for genuine belief motivation is necessary. Genuine belief then would be the criterion for normality. But genuine belief can be held when motivation fails. Cases of listlessness exemplify this possibility. There are also more fleeting and benign cases of listlessness. If I I'm very tired one day my motivations may become subdued, and though I may admit to having an obligation may feel no motivation to meet it. My belief is genuine; it's the same belief I had yesterday, say. But I am not motivated to do what I think I ought.

Maybe listlessness is an abnormality. Listlessness is a motivational defect, we might say. It occurs in roughly two different kinds of cases. Psychological trauma can induce listlessness. In these cases, listlessness seems like a disease, much in the way jaundice is characterized as an disease. It might also occur in people who would be motivated to do what they thought was right were it not for external stresses that impede motivation. Work related stress might be an example of this. Lack of motivation from this kind of interference seems similar to viewing objects in poor lighting. The internalist might describe normal people as those who do not suffer from psychological malady and external stress.

Suppose the concept NORMAL here refers to those who are not suffering from psychological malady or external stress. The internalist might say that the connection between

moral belief and motivation is conceptual in these cases. As a thesis sensitive to the scope of $I1_w$, then, consider $I1_n$:

> $I1_n$: Belief that Φ-ing is wrong entails having some motivation to Φ in people who do not experience psychological malady or undue external stress.

$I1_n$ is at the heart of the debate between externalists and internalists. The reason why is because the question of whether it is true turns on whether the so-called 'amoralist' is a coherent psychological figure. There are two kinds of amoralism. On the one hand there is what Brink (1997) calls the 'unprincipled' amoralist, whose motivational incapacity is due to some psychological or physiological malady. Mele's listless depressive, for example, is an unprincipled amoralist. But there is also the 'principled' amoralist, one who suffers no such impairment, but who believes sincerely and doesn't care at all. The question, then, is whether the principled amoralist is possible.

There is no well agreed upon answer to this question. Controversy surrounding $I1_n$ derives from the difficulty of imagining a normal person under normal circumstances who is not motivated to do what he thinks he ought. We have to suppose one believing that Φ-ing would be wrong to do while not enduring any excessively trying circumstances or illness, who appreciates the practical reasons that moral considerations provides and simply doesn't care. On the other hand, there doesn't appear to be anything incoherent about this character either. There doesn't seem to be anything about believing wrong that necessarily prohibits the absence of any resistance to Φ-ing. It's just a strain to find an example of this.

The failure of intuition to provide some headway against this problem is evident in cross-talk that sometimes emerges between internalists and externalists. Externalists who appeal to the principled amoralist to serve as a counterexample to $I1_n$ are accused of begging the question against the internalist. Brink, for instance, holds that the amoralist is a coherent psychological figure, and a counterexample to $I1_n$. But Smith objects that "[Brink] uses a prejudicial interpretation on the amoralist's reliable use of moral terms. He assumes that the amoralist's reliable use is evidence of her mastery of those terms; assumes that being suitably motivated under the appropriate conditions is not a condition of mastery of moral terms" (1994:70).

But the objection cuts both ways. If we take motivation under the appropriate conditions to be necessary for mastery of moral terms we beg the question against the externalist. The point is appreciated by internalists and externalists alike. Dreier, for instance, who argues in a later work for an externalist model of motivation, asserts that "the disagreement between Internalists and Externalists runs deep, and it lingers even in the face of clever intuition pumps. This debate in metaethics might be at a standoff, each side standing fast on its intuitions. Standoffs of this sort in philosophy are depressing" (2000:620). This is exactly the standoff that $I1_n$ invites. The internalist insists that normal people under normal circumstances cannot token genuine moral belief without being appropriately motivated. But externalists don't share this intuition. What we need is an independent reason for thinking that $I1_n$ is either true or false. Simply describing a normal person who is not motivated to do what he thinks he ought is not sufficient. But neither is it enough to merely assert $I1_n$.

In the remaining pages of this essay, I take on the challenge of making headway in assessing the truth of $I1_n$. Without relying on appeals to the (im)plausibility of amoralism, I

examine two approaches that purport to give independent reason for either accepting $I1_n$ or rejecting it. I first examine a recent challenge to externalism and a recent response given by James Dreier (2000). I argue that even if the challenge can be met, externalism about motivation faces a different problem, one that reveals an essential shortcoming to the approach. Then I look at a recent internalist account of moral motivation proposed by Smith (1994). Smith argues that moral beliefs are motivating to the extent that one is motivated by what she has most reason to do. But this view won't do either, since the demands for practical rationality may diverge from what one believes is morally required. Finally, I develop a view of moral motivation according to a rational version of sentimentalism. Rational sentimentalism offers a way of construing the link between moral belief and motivation via endorsements of warrant for feeling guilt. The solution offers a way of understanding the rational influence of fittingness endorsement over feeling, which in turn bears on what we are motivated to do. What falls out of the account is an explanation for just why $I1_n$ is true.

3.3 Externalism and Motivation

Externalism states that there is nothing specifically about the content of one's belief that Φ-ing is morally wrong that bears on what one is motivated to do. The fact that we are motivated to do what's right is due, rather, to contingent psychological facts about us. 'Us' here refers to we who are moralists. There is no interesting tie between belief and motivation in those who insincerely believe, nor any bond necessarily between moral facts and motivation. Further, where we find psychological or physiological impairment we also find this connection severed. The externalist

claims that there is something characteristic of the average (normal) person's psychology that explains why she is motivated to do what she thinks she morally ought.

The view lends itself to a charitable view of the amoralist. Since moral motivation will derive ultimately from one's psychological makeup, there's no trouble imagining a perfectly normal person who is not at all motivated to do what he thinks he ought to do. This person simply has a different psychological makeup from the rest of us. There's nothing he fails to understand or believe sincerely.

The externalist will also appreciate the role of those factors that shape one's psychology, in particular one's moral education and other social influences. The norms we follow and internalize play a crucial role in what we are motivated to do. Given the diversity of norms across different cultures, different people should be motivated in different ways.

But here's a challenge to externalism. Smith has recently argued that the externalist faces a dilemma. Here are his own words on the subject:

> [The externalist will] insist that what explains the reliable connection between judgment and motivation is a motivational disposition I have in virtue of which I count as a good person...But what exactly is the content of my moral motivation?..The explanation is only as plausible as the claim that the good person is, at bottom, motivated to do what is right, where this is read *de dicto* and not *de re*, and that is surely a quite implausible claim. For commonsense tells us that if good people judge it right to be honest, or right to care for their children and friends and fellows, or right for people to get what they deserve, then they care non-derivatively about these things. (1994:74-75)

Smith thinks that the externalist has to understand the content of one's (the moralist's) moral motivation as a motivational disposition to do what is right, read *de dicto*. *De dicto* motivation is a desire with the following content: to do what is right. By the externalist's lights, only *de dicto* motivation would be able to explain the motivation-tracking phenomenon. For example, say that you believe that abortion is morally wrong but after careful and mature consideration of arguments for *pro* choice, your belief weakens. Eventually you change your mind about the issue, now believing that women have a right to abortion. At this point, you obtain a motivational disposition to speak out in favor of *pro* choice, to argue against those who disagree and a host of other things. Your belief about what it is morally right to do has changed and with it your motivations. But why? According to the externalist, this must be because of some tendency in your character to alter the way you behave according to what you believe. This, Smith says, is a motivation to do what is right, read *de dicto*. And it must be. If you had a desire to do what is right, read *de re*, then you would be motivated non-derivatively by the features of abortion that make it wrong to do. But if this were the case, you wouldn't be motivated to speak out against abortion once your belief about its moral status has changed, since you still care about things like the intrinsic right to life of the fetus. The motivation to do what is right, read *de re*, will not explain the tracking phenomenon.

However, neither will it do to describe the moral motivation of a good person in terms of a disposition to do what is right, read *de dicto*. Good people "care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not just one thing: doing what they

believe to be right, where this is read *de dicto* and not *de re*" (1994:75). Smith likens the suggestion to an objection Williams makes when discussing the kind of moral philosophy that emphasizes impartiality (1976). Williams asks us to consider the following example. Suppose you can rescue only one of two people, one of whom happens to be your wife. Being the good, impartial, moral philosopher you are, you think to yourself that it would be morally permissible in this case to save your wife. But, Williams objects, this would be 'one thought too many.' Take things from your wife's point of view. She can rightly expect you to be motivated simply by the thought that she is your wife, where no further content is necessary to supply you with the relevant moral motivation. Anything more suggests that you do not have direct concern for your wife, which is after all what is morally required of a spouse. Moral motivation, thus, is the disposition to do what is right, read *de re*. When read *de dicto*, motivation amounts to a kind of moral fetish.

So here's Smith's dilemma for the externalist. If externalism about motivation is correct, then the source of moral motivation must come from something external to the content of one's moral beliefs, i.e., in some motivational disposition to do the right thing. If we read this desire *de re* we can't explain the motivation-tracking phenomenon, since if I change my beliefs about what is right I will no longer be motivated accordingly. If, on the other hand, we read this desire *de dicto* we fail to describe the motivations of a good person. Morally good people are motivated non-derivatively by what is right. Either way, the externalist is impaled on a horn of a dilemma.

How forceful is this challenge? In some cases it's hard to tell. There are instances in which a motivation to do the right thing, read *de dicto*, perfectly well describes the character of a

moralist. Consider an example taken from Lillehammer (1997). Say someone believes that she should sacrifice everything that she has for the greater good, perhaps even her life. This radical belief is new, however; she formerly believed that morality did not demand much in the way of personal sacrifice. Say this person is motivated to sacrifice everything she has, where this is read *de dicto*. Does this person have a moral fetish? It doesn't seem so. It seems reasonable to at least forgive this person if she does not desire to sacrifice everything she has in an underived way.

Further, non-fetishistic, *de dicto* motivation is not limited to cases of change in one's fundamental values. For example, suppose that I am trying to stop eating meat because I now think that contributing to an institution which treats animals inhumanely is morally wrong (whereas formerly I did not). My motivation to resist eating meat is read *de dicto*; I do not have an underived desire not to eat meat. In fact, I have a desire, *de re*, to eat meat. Nonetheless, I resist it. Am I not doing something laudable? It seems that *de dicto* desire can play a very useful role in regulating the behavior of good people.

Further, d*e dicto* motivation may be a necessary component of moral character. For example, suppose a father discovers that his son is a criminal, a murderer in fact.[75] The father knows that if he does not go to the police that his son will get away with the terrible crimes he's committed. The father believes that he morally should go to the police with information about his son, even though he doesn't want to. What moves the father is his *de dicto* concern to do what is right, while he has a *de re* desire not to do what is right. But these concerns are the marks of a good person are they not? Wouldn't it be odd if the father was motivated to turn his son in because he wants justice and retribution? It seems that a father's concern for his son

75 Example also taken from Lillehammer (1997:192).

should impel him to resist doing what he believes he morally should do—and that when he does do the right thing, it's because he desires to do the right thing, read *de dicto*.

De dicto desire doesn't seem to be that bad. Consider now a positive externalist account of moral motivation that purports to meet Smith's challenge.

Dreier (2000) has recently offered a way of understanding how *de dicto* motivation can play a role in the psychological framework of the moralist. He makes use of a notion called a maieutic end (Schmitz 1994). Like any end, a maieutic end is a goal or aim. But a maieutic end is different from other kinds of ends. Think of the familiar notion of a final end. Think of getting exercise. One may desire to get exercise for its own sake, that is, just to get it, or just because one values exercise. On the other hand, one may desire exercise for the benefits that it provides to one's health. In this case, the goal of exercising is instrumental to the aim of being healthy or benefiting one's health. Alternatively, one may have the goal of running several miles to get exercise. Here, running is not exactly instrumental to the getting of exercise, but is rather constitutive of it. Unlike final, instrumental or constitutive ends, a maieutic end has a second-order quality—something that's achieved by coming to have other ends. Here's an example from Schmidtz:

> Suppose that for Kate, becoming a surgeon is an end. Perhaps it is an end because Kate thinks becoming a surgeon will be prestigious, in which case becoming a surgeon is an instrumental end. Kate becomes a surgeon in order to do something else, namely, to secure prestige. But maybe for Kate becoming a surgeon is an end in itself. How could a career in medicine come to be a final end?

> Maybe it happened like this. When Kate was a teenager, she had no idea what she wanted to do with her life, but she knew she wanted to do something. She wanted goals to pursue. In particular, she wanted to settle on a career and thus on the goal or set of goals that a career represents. At some point, she concluded that going to medical school and becoming a surgeon would give her the career she wanted. So she went to school to pursue a career in medicine. She has various reasons to pursue this goal, of course, but she also pursues it as an end in itself.... (1994:228)

We might imagine Kate developing even further ends, like relieving suffering or being a respected member of the medical community. These further ends she might value for their own sakes. Kate's coming to settle on a career would then be achieved by coming to have these other final ends. Her desire to settle on a career leads her to decide to become a doctor, which leads her to value the things that doctors value intrinsically.[76]

A maieutic end is much like a second order desire. Kate's end to have a rewarding career is just the desire to have one, a desire that's achieved by coming to have other desires. Kate wants to have other desires (to relieve suffering, etc.), since these constitute having a rewarding career in medicine. Suppose, now, that one has the second order desire to do what is right. This desire is one in which one desires to have other desires, namely, those to do acts that are right. Putting the matter another way, we might say that a second order desire is a desire to do what's right, read *de dicto*, but also a desire to have other desires, read *de re*. For example, I am a meat

76Thus, Schmitdz's suggestion offers us a way of assessing ends, which, according to traditional accounts of rationality, cannot be done. Consider the words of Bertrand Russell: "'Reason' has a perfectly clear and precise meaing. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends" (1954:8).

eater but I believe that eating meat is wrong.[77] Because I am a meat eater, I desire to eat meat. However, I also have a desire to do the right thing, read *de dicto*. Because of this desire, I routinely resist eating meat and avoid situations in which I might be tempted to eat meat. As well, I remind myself of the inhumane practices of the meat industry. Over time, I lose the desire to eat meat and gain the desire not to eat meat, read *de re*. My desire to do what's right is thus a second order desire which causes me to have a first order desires to do what I think is right. Call this the 'maieutic desire' model of moral motivation.

The second order desire to do the right thing operative in the maieutic desire model ensures that motivation will reliably track changes in belief about what is morally right. Because I desire to do the right thing, read *de dicto*, I will be motivated to do those things that I think are right, regardless of what they are. Over time, I will develop first order desires to do those things that I believe are right.

The maieutic desire model also anticipates that changes in our motivations will happen gradually. When we change our beliefs about what we should do, some tension is likely to develop between our present motivations and the motivations we think we should have. We should expect some tension between what we are motivated to do, where this is read *de re*, and what we are motivated to do, read *de dicto*. For example, when I come to believe that eating meat is wrong, there is both conflict between my desire to eat meat and my belief that I shouldn't, and tension between my desire to eat meat and my desire to do the right thing, read *de dicto*. I desire to have a different desire, namely, the desire not to eat meat. But I presently don't. It's only after taking measures to have this desire that I obtain it. In as much as my desire to do the right thing is effective, I will take measures to ensure that I obtain the corresponding *de*

77 I'm a vegetarian actually.

re desire to do the right thing, since when I change my values, I very likely do not have the *de re* concern to do what I now believe I should.

The maieutic desire model can also accommodate cases in which belief and desire come apart. On the one hand, second order desires are not always effective. One may believe that one should do what one presently does not want to do without thereby being motivated, in the end, to take measures to be motivated accordingly. I may believe that eating meat is morally wrong but feel nothing in the way of refraining from going to the after school luncheon where hamburgers are regularly served, or making my beliefs public to my family so that they might offer vegetarian dishes at Sunday dinners. Our motivation to do what we think is right, whatever that happens to be, is clearly not always as strong as the motivations we currently have (so long as they are different of course). But they are occasionally, if not often. Thus, it is only if the second order desire is strong enough to outweigh competing first order desires that one's first order desires will reliably track changes in one's moral beliefs.

On the other hand, the model can also accommodate cases of listlessness. Say that I believe that eating meat is morally wrong but don't have the slightest inclination to resist eating it. What's going on here? The externalist can say that I want myself to have a concern for the suffering of animals and the injustices of the factory farming industry, concerns that would give me pause when thinking about what to order at my favorite restaurant. Unfortunately, however, I don't have any such concerns. My desire to do the right thing in this case, read *de dicto*, is wholly ineffective in generating a desire to resist eating meat, read *de re*. I want to be motivated for the right reasons to do what I think is right, but I am just not motivated at all to do what I believe is right.

The model can do all of this, purportedly, with a second-order desire that is not fetishistic. Once I accustom myself to doing those things that I think are right, I obtain desires to do those things, read *de re*. Further, this is exactly what I want in my second order desire. I want to be motivated by those things that I believe are right to do. I will be motivated, so long as I take measures to do those things that I believe are right to do. My second order desire to do the right thing will impel me to take measures to ensure that I desire to do those things that are right. And eventually, what motivates me to do them is a concern directly for their right-making features.

Still, it might be thought that the moralist's first order desires are infected with fetishism (Smith 1997). On the maieutic desire model, one has the desire to have other desires that would contribute to one's doing what one thinks one ought to do. But those other desires are arrived at only instrumentally, that is, as a means to be motivated to do what one thinks one ought. A good agent, on the other hand, would care noninstrumentally about those things that his motivations bear upon. For example, say I have a second order desire to have other desires to do things that I think are right. Since I believe that eating meat is wrong, I have a desire to desire not to eat meat. Smith claims that my desire not to eat meat comes about only as a means to satisfy my second order desire—and that this is not how they should come about. If I were a true moralist, I should be concerned directly with the suffering of animals and so forth, not with achieving the end of desiring to do what is right (*de dicto*). As he puts it, "[as the second order desire model] describes morally virtuous people they are ultimately motivated by the fact that they have noninstrumental desires to perform acts with right-making features, not by the fact that they have

noninstrumental desires to perform acts with the features that they believe to be the right-making features themselves" (1997:115).

Dreier denies this. When I change my beliefs about the practice of eating meat, I establish a motivation not to eat meat only if my second order desire, the desire to have the desire not to eat meat, is stronger than my current desire to eat meat. Thus, my first order motivations change only if my desire to do what is right, read *de dicto*, is stronger than them. I will resist eating meat only if I really want to do the right thing, whatever it happens to be. And my second order desire is not a concern for the right-making features of vegetarianism, but is just a concern to do the right thing. However, this does not mean that I have a moral fetish. It's true that my desire not to eat meat derives from my second order desire to have desires to do things that are right. But this ultimacy, says Dreier, is only causal. My desire to do the right thing, if effective, will motivate me to take measures to ensure that I obtain, over time, the desire not to eat meat. Once the desire not to eat meat is in place, however, it no longer depends in any way on my desire to do the right thing.

But what about changes in the belief that something would be morally permissible to do? As it turns out, even if Dreier is right, the externalist faces a new dilemma. Either Smith is right and first-order concerns for the things that are right to do are merely instrumental in satisfying our second-order concerns to have the first-order ones, or Dreier succeeds in explaining how second-order desires can generate first-order desires in a way that is non-fetishistic. If Smith is right, then fetishism infects our first-order desires. If Dreier is right, however, we are left with no explanation for why our motivations track changes in belief about what is morally permissible to do. Let me explain.

Believing morally right expresses 'believing that one is morally obligated to Φ.' Believing morally permissible expresses 'believing that it is okay to Φ.' Here's the difference between these two phrases. Say I used to believe that eating meat is morally permissible while now I believe that it is morally wrong. My former belief is 'believing permissible.' Now I believe that I ought not to eat meat. My belief about what I ought to do is 'believing right.' When I change my belief about eating meat I change my belief about what I ought to do. Since the belief that I arrive at is one about what I am morally obliged to do, call this is a change in 'believing right.' Now consider the reverse. Say that formerly I believed that eating meat was morally wrong to do and that now I no longer believe that it is. In this case, I have changed my belief about what is morally permissible to do. Since the belief I arrive at is one about what I am morally allowed to do, call this a change in 'believing permissible.' Dreier's model may be able to explain why motivation tracks changes in believing right, and so answer Smith's challenge to the externalist. What he cannot explain is why motivation would track changes in believing permissible.

Say I formerly believed that eating meat was morally wrong but have now come to believe that it's okay to eat meat. I've become convinced that animals in fact do not suffer, and so do not have any right not to be treated inhumanely or killed. Given my former convictions about the the matter, I presently have a concern for the way animals are treated by the factory farming industry. If changes in moral belief reliably elicit changes in motivation, I should lose this concern over time. But what is to cause me to lose this concern? According to the externalist, it cannot be my current belief that eating meat is morally permissible, since there's nothing about the content of that belief that should elicit such a change.

The maieutic desire that Dreier makes use of may explain why our *de re* concerns would change over time with a change in our beliefs about what is morally wrong to do. With our moral habits firmly in place, something is required to force a change in these habits once we come to believe that we should have different ones. Being predisposed to be concerned for the right making features of an action puts pressure on us to have those concerns, whatever they happen to be—where we take measures to remind ourselves of the fact that what we formerly believed is incorrect, and that we should now not do nor condone the action that our belief was about. However, there's an asymmetry between a second order desire to do the right thing and a desire that causes our *de re* concerns to change in line with a change in our beliefs about what is morally permissible to do.

The problem is an insensitivity to grounds we have for thinking that Φ-ing is morally wrong. Where these grounds disappear, there's nothing to explain why our motivation to resist Φ-ing would also disappear. This suggests that the content of our moral beliefs really do bear on what we are motivated to do. It is no small advantage to an internalist account of moral motivation that it can explain why motivation would track changes in moral belief in both directions. By linking motivation to the content of moral belief, changes in moral belief, in whatever direction, will reliably produce changes in motivation. The fact that the motivation-tracking phenomenon is bidirectional shows that a sensitivity to grounds is an essential part of solving the problem of moral motivation.

3.4 Rationalism and Motivation

Rationalism is the view that there is some rational connection between believing right and being motivated to do what one believes is right. It's an internalism about motivation in that it asserts an analytic connection between belief and motivation, but it's different from internalisms that take this connection to be conceptual, where sincere belief must involve motivation. The rationalist has no problem with the idea that people can believe sincerely that something would be wrong to do yet not be inclined to resist doing it. It's just that these people are irrational in some sense.

Say that Tommy wants to attend a lecture downtown at 3:00. In fact, he thinks he should go. (Say his major professor will also be there). In order to do this, he will have to take the 2:00 subway, and in order to do that, he will have to leave his apartment now and walk seven blocks. But the thought of walking seven blocks irritates Tommy. He doesn't like to walk even relatively short distances. Seven blocks, though not long really, is not a short distance. Tommy has no motivation whatsoever to walk to the subway station. He thinks he should go to the lecture, but that consideration fails to get him to leave his apartment and walk to the subway station. So he stays home.

There's something wrong with Tommy. He isn't motivated to do what it would take to do what he thinks he should. Granted, it's not particularly uncommon for people to decline doing what they think they should. But Tommy is not motivated at all to do what is necessary to achieve what he thinks he should, and this is a little strange. The thought that he ought to Φ doesn't motivate him to do what is necessary to Φ. Yet, the consideration that one should do

something typically motivates one to do what it takes to do it. We might say that Tommy would be motivated to walk the seven blocks to the subway station if he was 'in his right mind' or something like that. It seems that, if we were to talk him into making the trip, we would merely be talking him into doing what he wanted, in the end, to do anyway.

Korsgaard put the matter this way:

> Being motivated by the consideration that an action is a means to a desirable end is something beyond merely reflecting on that fact. The motive force attached to the end must be transmitted to the means in order for this to be a consideration that sets the human body in motion—and only if this is a consideration that sets the human body in motion can we say that reason has an influence on action. A practically rational person is not merely capable of performing certain rational mental operations, but capable also of transmitting motive force, so to speak, along the paths laid out by those operations. (1986:13)

There is nothing wrong with Tommy's ability to reason about what he should do or how to achieve what he thinks he should do. In so far as he can perform certain 'rational mental operations' he is normal, or rational. What's odd about him is that his reasoned belief about what he should do doesn't motivate him to do it. His behavior is not what we would expect from one who believes that Φ-ing is what one has most reason to do. And to be lacking in this motivational capacity is odd, since the whole point in deliberating about what to do is to actually do what one thinks one should.

This is the making of a different kind of internalism. It is this: claims about what we should do, or practical-reason claims, will motivate persons in so far as they are practically rational. Tommy is not motivated to do something that he thinks he should. He's irrational in the sense that the motive force of what he thinks he ought to do is lacking. Applying the case generally, reasoned belief will elicit motivation in so far as one is moved by what one believes one has most reason to do. This is the central thesis of rationalism or R:

R: Belief that Φ-ing is right entails that one will be motivated to Φ in so far as one is practically rational.

As stated, however, R is vacuous. We want to be motivated to do what we believe we have most reason to do, but why? Why should we trust our reasoned beliefs? What's so good about being practically rational? Answers to these questions require a substantive concept of practical rationality.

According to Korsgaard, being practically rational means both having the ability to reason correctly about things and being responsive to beliefs about what it is best to do. Both parts are important. The motivation part connects belief to motivation. Now turn to the assertion that one be able to perform certain 'mental operations.' The idea seems to be something along the lines of 'being best able to discern the best reasons for Φ-ing.' If someone is able to do this, then she will act on those reasons, in so far as she is practically rational, and do what she has most reason to do.

What we get is a concept of rationality along the lines of what Williams (1980) thought were conditions for someone having a good reason to Φ. Call these conditions for 'full' rationality. Williams suggests three. The first is that an agent has no false beliefs. Say, for example, that Sally thinks she sees a black snake in her closet, but that what she sees is really a pile of black socks. As far as she can tell, she has a good reason to close the door immediately and leave the room. But in fact she has no such reason. If she were fully rational, however, she would recognize this, and act in ways consistent with what she correctly sees as a reason to calmly pick up her socks and put them in the sock drawer.

The second is that an agent have all relevant true beliefs. An agent "may be ignorant of some fact such that if he did know it he would, in virtue of some element...[his motivational set]..., be disposed to Φ: we can say that he has a reason to Φ, though he does not know it" (Williams, 1980:103). Recall Tommy's desire to attend a lecture downtown. Suppose that he knows only of the subway that will get him there in time but doesn't want to walk the seven blocks to get to the subway station. In fact, however, there is a bus that makes a stop only two blocks away, and is scheduled to arrive at the lecture hall even a few minutes before the subway. But Tommy doesn't know this. According to the second condition for full rationality, Tommy would know this if he were fully rational. This is a fact relevant to his deliberation in the sense that, given the motivations he already has, it would influence his decision about what to do.

The first two conditions may seem controversial. What Tommy and Sally need is not more rationality; they need more information. The point, however, is that our reasoned beliefs are made on considerations that bear on whether we have all of the relevant information. When we arrive at conviction about what is the best thing to do, we generally do so under the

assumption that we do have all of the relevant information. If Sally believes that she has most reason to close her closet door, it's plausible to assume that she also thinks there's something dangerous in her closet. She thinks she should close the door, but if she were made aware of the fact that there was nothing to be afraid of she likely wouldn't. It's possible to believe that one has most reason to Φ while admitting that one is uncertain about all of the relevant facts, of course. I might believe that I have most reason to go to school today and admit that I could land in a car accident on the way. But assuming that such things will not happen, I believe I have most reason to go to school today. If I were made aware of the fact that I would land in a car accident on the way to school, I would no longer believe that I have most reason to go to school today (hopefully).

Third, the fully rational agent deliberates correctly. Practical reason has most often been construed, as Korsgaard herself does, as a means-end deliberative process, where one has a desire to Φ and deliberates about what means would be most efficient to Φ given other desires or preferences one has. But there are other ways one might deliberate about what to do. Some have emphasized the role of imagination (Williams 1980, Lewis 1989). This is to bear in mind possible action, situation or choice-scenarios, to weigh them against one's responses and beliefs and to decide from there what to do or how to think or feel. There are also the more systematic efforts we make to find 'reflective equilibrium' among our specific and general evaluative beliefs (Rawls 1971). In moral argument, for example, we generally start at a point of disagreement. Then we proceed to establish matters—themselves relatively uncontroversial—about which we do agree. We attempt to find principles that would explain why we think those things about which we do agree are true, all in the hope of coming to some agreed principles that would

support one's original view over another. But if we derive from that principle untoward or unwanted conclusions we would then have reason to reject the principle, and so on. We keep reasoning in this way until we find a sort of balance between our general normative principles and specific normative beliefs.

Much has been made of this last form of rational deliberation. Smith (1994) has emphasized the efforts to achieve reflective equilibrium among one's motivations—that this is what a practically rational person will do. This is plausible. A set of beliefs achieved by reflective equilibrium exhibits coherence and unity. Such a set consists of beliefs that are mutually supportive and without inconsistency. If these are beliefs about what to do, and if what is believed to be done reliably influences a practically rational person, then her motivations will exhibit this same coherence and unity. Achieving reflective equilibrium among one's beliefs about what to do will also establish coherence and unity among one's motivations.

However, it could be the case that having all of the relevant information about what to do does not settle the matter. One could know all of the consequences of one's actions, be perfectly aware of all the detail of the circumstances in which one acts and so forth, without having what is needed to settle the matter of what to do. It could be the case, in other words, that 'what to do' is not itself an objective matter, that it cannot be discerned just by knowing all of the relevant natural facts. Two people may disagree about what to do despite each knowing all of the relevant facts that pertain to what to do. If 'what to do' was a relative issue, rational agents would not necessarily agree in principle.

Nonetheless, the fully rational agent is in the best position to decide what to do. Even if practical matters were relative, the fully rational agent wouldn't provide bad advice. There may

be another fully rational agent offering different advice, but she's not wrong either. Perhaps an error theory is true, and everyone's moral judgments are false. Still, the fully rational agent's assessments are no worse than anyone else's.

Further, the advice of the fully rational agent is agent relative. Consider the difference between what a fully rational agent would do in circumstances C, and what we should do in C. Smith illustrates with the following example:

> Suppose I have just been defeated in a game of squash. The defeat has been so humiliating that, out of anger and frustration, I am consumed with a desire to smash my opponent in the face with my racket. But if I were fully rational, we will suppose, I wouldn't have any such desire at all. My desire to smash him in the face is wholly and solely the product of anger and frustration, something we can rightly imagine away when we imagine me in my cool and calm fully rational state. The consideration that would motivate me if I were fully rational is rather that I could show good sportsmanship by striding right over and shaking my opponent by the hand. In this case, does it follow that what I have reason to do *in my uncalm and uncool state* is stride right over and shake him by the hand? (1995:111, original italics)

I do not have any such reason. What I have reason to do in this situation is:

> smile politely and leave the scene as soon as possible. For this is something that I can get myself to do and it will allow me to control my feelings. Moreover...*this is exactly*

> *what my fully rational self would want my less than fully rational self to do in the circumstances that my less than fully rational self finds himself....* [I]t is not something I would be motivated to do if I were fully rational because it is not something that I would have any *need* to be motivated to do if I were fully rational. (1995:111, original italics)

Φ-ing isn't right to do just in case a fully rational agent would be motivated to Φ, since she doesn't necessarily have anything in common with her less than fully rational self in her circumstances. Rather, the sense in which one should want to do what one believes is right is that in which one's fully rational self is sensitive to the circumstances one actually finds oneself in. We shouldn't do or be motivated to do what our rational selves would do, but what our rational selves would advise us to do.[78]

With a substantive concept of practical rationality, turn now to how R might be interpreted as a claim about motivation. Like the more general internalist thesis I1, R can be interpreted either strongly or weakly. On a strong reading, one is motivated, in the end, to do what one has most reason to do in so far as one is practically rational—where the phrase 'practically rational' means believing that one's fully rational self is advising one to Φ in a particular set of circumstances. Thus R_s:

78 It is thus clear why Smith emphasizes the role of achieving reflective equilibrium in our mortivational set. This effort requires reasoning about matters, getting clear on the details of the circumstances one finds oneself in, including facts about oneself. The reasoning that establishes beliefs about what to do establishes also the motivations we are to have. These motivations are, according to us anyway, the motivations we should have; they are the motivations we think our fully rational selves would advise us to have. Thus, the belief that Φ is right to do in circumstances, C, should cohere with beliefs that we have. If it coheres well with other beliefs that we have, if it plays a role in establishing reflective equilibrium among our many beliefs, then we rationally should Φ. Having the motivation to Φ, then, is underwritten by the same demands of coherence and unity that we have for our beliefs about what to do.

R_s: Belief that Φ-ing is right to do in circumstances C, entails that one will be motivated to Φ, in the end, in C, in so far as one is practically rational.

R_s has it that practically rational people are those who have an overriding motivation to do what they think is right. Likewise, if one does not have an overriding motivation to do what one thinks one should, one is practically irrational. This division is natural, I think. If I believe that I ought to Φ and am not motivated in the end to Φ, then I have failed to move myself to Φ; I am practically irrational at least in this sense.

As a weak thesis, R says that one is practically rational to the extent that one has some motivation to do what one thinks one ought or R_w:

R_w: belief that Φ-ing is right to do in circumstances C entails having some motivation to Φ in C in so far as one is practically rational.

This way of carving things up also captures something intuitive about motivation. The odd cases of moral motivation involve subjects who experience no motivation whatsoever to do what they think they ought. The listless person simply doesn't care at all about doing what he thinks he should, while the amoralist strikes us as bizarre and extremely wicked. One who doesn't do what one thinks one ought and feels guilty about it, on the other hand, is all too commonplace. Demarcating the line between practical rationality and irrationality by these cases appears to follow this intuition.

Both versions have an easy time with cases of listlessness and the like. Anyone who fails to be motivated in the end to do what they think they should is practically irrational. It applies to those who have some motivation but not enough, as well as to those with none at all. The depressed person would do what he thinks he should were he practically rational. He believes that Φ-ing is the best thing to do but isn't motivated because he's depressed. Still, he would if he wasn't.

Both also anticipate that changes in motivation track changes in belief about what is permissible to do. If, for example, I was formerly a vegetarian but have since become convinced that there is nothing wrong with eating meat, then according to the strong version, I should lose any apprehension I have towards eating meat. That is to say, my fully rational self would advise me not to have this motivation, since I now believe that there are no grounds for believing that eating meat is wrong. Any lingering resistance is a sign of my practical irrationality.

However, the weaker version expresses a sensitivity to reasons that bears out a significant advantage over the strong version. First of all, the weak version can explain why we would feel conflicted in cases where reasons compete for determining whether Φ-ing is best. For example, say I am practically rational and that I discern both good reasons to go to the store and good reasons not to go. In the end, the good *pro* reasons appear to outweigh the good *con* reasons, so I believe I should go to the store. According to R_w, I should have motivation to both go to the store and not go to the store, causing me to be conflicted about what to do. But this is how it should be. I have reasons *pro* and *con* here. It seems that my fully rational self would appreciate the conflict of reasons in my deliberations about what to do.

Further, the strong version undermines the difference between one who does not have an overriding motivation to do what's right and one who has no motivation to do what's right. According to the strong version, one is practically irrational to the extent that one does not have an overriding motivation to do what's right. Thus, one who has no motivation to do what's right is just as practically irrational as one who does not have an overriding motivation to do what's right but is conflicted about what to do. And this does not seem right. If I have reasons *pro* and *con* for going to the store I should be conflicted about what to do. Were I not to be conflicted in this situation I would be more practically irrational than were I simply to fail to appreciate when the *pro* reasons outweigh the *con* reasons. That is to say, it is more odd to ignore completely the force of *con* reasons than to be compelled by them but not enough so that I act on them. The weak version accommodates this difference.

This an advantage that becomes even clearer when we think about the gradual change that we experience when our beliefs about what to do change. In these cases, our fully rational selves are advising us to have motivations that we currently don't have (or advising us not to have motivations we do). Both strong and weak versions entail that we are practically irrational when our preceived bases for Φ-ing disappear. But the weak version would predict a rational hesitancy if these bases do not disappear altogether. If I come to believe that Φ-ing is no longer the best thing to do, but that there is still some reason to Φ, then my resistance to change is rational. The strong version denies this. On the strong view, a lingering motivation to Φ does not make sense if one does not believe that Φ-ing is best to do, regardless of any countervailing considerations. Thus, the strong view implies that gradual change of motivation is always irrational. And this, I think, overstates the connection between reasons and motivation.[79]

79 This argument applies *mutatis mutandis* for gradual change in belief about what is permissible.

R_w is the best version of rationalism. However, problems with this position remain. In the next section, I develop a critique of rationalism. Though there is something irrational about believing that one ought, all things considered, to Φ and having no motivation to Φ, there is nothing necessarily irrational about believing that Φ would be morally right to do and having no motivation to Φ.

3.4.1 Critique of Rationalism

Rationalism is a view about what we have most reason to do. Our fully rational selves are privy to all of the relevant facts, and deliberate correctly. They give us advice about what it is best for us to do in our particular circumstances and about what motivations we should have. But moral beliefs are different from beliefs about what is best to do. It's possible to believe that Φ-ing is morally right and yet believe that Φ-ing is not best to do. Can rationalism be a motivational thesis about our moral beliefs?

For the rationalist, motivation is tied to what our fully rational selves would advise us to be motivated to do in some set of circumstances. In order for rationalism to succeed in explaining the motivating quality of moral belief, it must be true that every time we believe that Φ-ing is morally right to do, we believe that Φ-ing is the best thing to do. The rationalist is thus committed to a view about belief, which we might describe in the following way:

> Rb: to think that Φ-ing is morally right to do in circumstances C is to think that Φ-ing is the best thing to do in circumstances C.

The phrase 'morally right' would be read as 'not morally wrong.' Motivation has a special tie to our perceived moral obligations, not to what we deem morally permissible or to those things we think go above and beyond the call of duty. But is it true that perceived moral obligations amount to a commitment to what is best?

According to one theory of rationality, what one has most justifying reason to do is what will serve one's best interest. It is irrational to do what one believes will be worse for oneself. Call this the 'self-interest' theory of rationality.[80] The self-interest theory of rationality insists that the only reasons that can bear on what to do are those that speak to what's in one's best interest. It also seems reasonable to allow, however, that other kinds of reasons could bear on what to do all things considered.[81] Moral reasons, in particular, also seem like perfectly good reasons for action. Suppose, for example, that the butler stands to gain a small sum of money if the lord of the manor dies. The butler could easily kill the lord of the manor but this would be morally wrong to do. The fact that it would be morally wrong to do outweighs what he stands to gain by doing so. The butler could decide, correctly, that killing the lord of the manor is not the thing to do. Thus, Φ-ing can be best to do because Φ-ing is morally right to do. Since it is rational to do what is best, it is sometimes rational to do what is morally right. Perhaps acting in

80 See Parfit (1984).

81 Moral and self-interested considerations occupy distinct normative spaces. Not everyone has believed this, but the idea is more well accepted than not. Consider Parfit's remarks on the subject:

> A moral theory asks, not 'What is rational?, but 'What is right?' Sidgwick thought that these two questions were, in the end, the same, since they were both about what we had most reason to do...A century later, these two questions seem further apart. We have expelled Egoism from Ethics, and we now doubt that acting morally is 'required by Reason'...There are many cases where it would be better for someone if he acted wrongly. In such cases we must decide what to do (1984:129).

one's best interest is not always the best thing to do. Acting rationally is sometimes acting morally.

Thus, there are at least two different kinds of considerations that can bear on what it is best to do: that Φ-ing would be morally right to do, and that Φ-ing would be in one's best interest to do. In as much as moral reasons for action are genuine reasons, they could outweigh self-interested reasons for action. And in as much as self-interested reasons for action are genuine reasons, they could outweigh moral reasons for action. Each kind of reason appears to have its own normative force, and can play a strong enough role in normative deliberation to determine what it is best to do.[82]

However, though moral reasons can themselves justify Φ-ing, the force they have must derive from considerations that speak to ultimately what is in one's best interest. That is to say, we cannot answer the question 'why be moral?' by saying 'just because.' Moral reasons can justify Φ-ing but they are not self-justifying. Reasons that speak to what is one's best interest, on the other hand, are. Consider how strange it is to ask 'why be concerned for what is in my best interest?' It's like asking how considerations that bear on what is good for oneself could benefit oneself. We might imagine one saying, 'why should I pay attention to what might contribute to my own end? What's in it for me?'

82 The pluralistic view of reasons also leaves room for controversial cases in which it is not clear what to do. Consider one who is considering cheating on his partner. Suppose that it is a one-time affair, and that it could be done discreetly. Suppose also that this person is unhappy in his relationship and wants some excitement and passion. However, he would feel guilty if he actually did what he wanted to do. Would it be rational for this person to go through with it? The fact that he would be doing something wrong certainly speaks against the infidelity, while the fact that it would satisfy his desires speaks comparably in favor of it. It's also plausible to argue that this person wouldn't be doing something that wrong. It's not as wrong as what Judah does anyway. On the other hand, the gain he can expect is not that much either. He might be able to satisfy a desire for excitement, but the feeling would likely be fleeting and, in the end, little to remember. What to do, I submit, is not clear in this case.

Many have seen the need for an external justification for morality. Not surprisingly, the consensus is that moral reasons override self-interested ones because it is, in the end, in one's best interest to be moral. The suggestion here is that being disposed to act morally is rational because it is in one's best interest (Gauthier 1986, Frank 1988). Suppose, for example, two people can engage in a potentially profitable venture. Their potential for gain arises from the natural advantages inherent in the division and specialization of labor. So together they have the necessary skills to launch a successful venture—working alone, their potential is much more limited. Suppose now that each will have opportunities to cheat without getting caught. If only one of them cheats, he does very well. If the other also cheats, he, too will not get caught and will do better than by not cheating. But if both cheat they both do worse than if both had been honest. A straightforward way of ensuring that both do well is that both be disposed to act honestly and fairly. If such dispositions are detectable, then others will have signs that one is a worthy companion in a cooperative venture.

Of course, it's in the best interest of each to fake a moral disposition and then cheat when he can get away with it. But these are very difficult things to do. At least for most of us, who are not good actors, giving the impression of honesty when one is really not is hard. We may be able to lie effectively every now and then but to maintain a veneer of honesty for a long period of time is difficult. As well, even if we could fake our way into cooperative situations, it is difficult to know when one really has a good opportunity to cheat. We would have to be very clever and conservative in our criminal activity to reap the rewards. One bad step and we not only lose what we may have gained by succeeding but would be likely displaced from the cooperative situation by our partners. For most of us, then, simply being disposed to honest and

fair behavior is the best policy. We won't be able to gain from situations in which we might have taken advantage of others but we don't lose by being excluded from the cooperative situation altogether. Furthermore, we don't have to extend any effort into calculating the odds of getting caught or how to respond to a situation in a way that an honest person would. Just being a good person is simpler and effective. We can reliably gain from what opportunities cooperation offers.

All of this, however, true though it may be, doesn't show that acting immorally is always irrational. It's rational to adopt a policy of good behavior but it does not follow that it is rational to act morally all of the time. Like all policies, a policy of good behavior is defeasible. Adopting a policy to remain honest and fair may be in one's long-term best interest but it would likely (indeed, inevitably) incur some cost. If the costs are great, one would be irrational to take them on.

Rb is therefore false. Accordingly, combining Rb with the strong and weak versions of rationalism will yield spurious claims. Let's look at these now. On the strong version, moral belief would entail having overriding motivation in so far as one is rational, or Rm_s:

> Rm_s: Belief that Φ-ing is morally right in circumstances C entails having an overriding motivation to Φ in C in so far as one is practically rational.

Rm_s is true if our fully rational selves would recommend that we have an overriding motivation to Φ in our particular circumstances if Φ-ing is the morally right thing to do.

But our fully rational selves may not advise us to be motivated to do what is morally right. We can appreciate this with an example. Consider this. Judah has been having an affair.[83] He is a wealthy man. He has built a prosperous life for himself. His family is healthy and comfortable, and he is well respected and liked. But the affair threatens everything that Judah has. His mistress, jealous of the life he keeps from her, has threatened to reveal the affair to Judah's wife as well as other financial indiscretions that he has confided to her in private moments. If these were to become public, Judah would be ruined. Desperate to close the relationship, Judah pleads with his mistress to leave quietly, but she will not agree to it. Her threats persist, and she demands that he leave his wife and marry her. Convinced that his mistress is incorrigible, Judah begins to despair. A friend then offers some advice to Judah: have the mistress killed. The job can be done quietly and without any risk to Judah himself. All it will take is some money, discretion, and the will to go through with it. Appalled at the suggestion Judah initially refuses. But he cannot allow his mistress to ruin his life. After several days Judah reconsiders and has her killed.

The case is controversial. It's tests our intuitions about what reason Judah has to kill his mistress. The fact, so he thinks, that it would be morally wrong provides him with a reason not to go through with what he thinks he must (lest he be ruined). But the likelihood that his mistress will reveal the affair also gives him some reason to do it. What counts here are the reasons Judah has for acting, and, given the decision that he makes, whether acting in the way that he does is acting on the best reasons he has.[84] The question elicited by this example is

83 Example taken from Woody Allen's *Crimes and Misdemeanors.*

84 Two points may be useful to mention here. First, there is a difference, recall, between explanatory reasons and justificatory ones. Explanatory reasons describe an agent's motive for acting. One may decide to get some ice cream, for example, because of a craving for ice cream. This desire explains one's action. This is one's reason for acting in the sense that it is why one acted. Justifying reasons justify an action (if they're good ones). If by getting some ice cream one thereby satisfies the wish of a child, one might be justified in doing that. Both

therefore whether Judah is rational in the sense of performing certain 'mental operations.' He doesn't have any sort of motivational defect. He is not one who thinks he ought not to kill his mistress but doesn't care. If this man is practically irrational, he's so not for the same reason that Tommy is practically irrational. If Judah is practically irrational, it's because he comes to the wrong decision about what to do, not because he isn't motivated to do what he thinks he ought.

The point can be pushed further by considering people who are good cheaters and care little for acting morally. A clever sociopath, for instance, may be able to detect those opportunities he really has for cheating and be able to take advantage of them. Since he would suffer none of the emotional cost of wrongdoing, there's no reason to advise him to resist doing what would be best for him but not so much for others. A sociopath's fully rational self would advise him to adopt a policy which recommends acting morally when the chances are he will be caught, but acting immorally if the chances are that he will not. If he can also fake a moral disposition, this policy would be best for him in the end. Thus, the rationalist would have it that the analytic connection between moral belief and motivation for most of us is the same as that between immoral belief and motivation for some others. While it's rational for most of us to adopt a policy of good behavior it's rational for others to adopt the opposite policy. Believing that Φ is morally right to do, thus, not only has no connection to motivation in these other folks,

---

justifying reasons and explanatory reasons are reasons in the sense that they render an agent's action intelligible. But while explanatory reasons speak to why the agent acts, justifying reasons indicate some prescriptive requirement to act. Explanatory reasons cite psychological states of the agent, while justifying ones point to a prescription. As such, justifying reasons are seldom the motivation one has to act. Judah's justifying reason for killing his mistress is not his desire to kill her, but the irreparable harm that will befall him if he doesn't. His desire to preserve himself, his work, and his family, on the other hand, explains why he does what he does. Judah also has good reason not to kill his mistress, that it would be morally wrong to do. But, obviously, this doesn't explain why he did what he did. The kinds of reasons that pertain to the present line of inquiry are justifying ones.

Second, neither explanatory nor justifying reasons need be good ones. All that matters is what they purport to do. Citing the motivations of an agent may explain her action but they may not. Reasons can be explanatory but fail to explain an agent's action, for example. Same goes for justifying reasons.

rather, believing that Φ is in one's best interest does. Given their cleverness and lack of moral sentiment, their fully rational selves would have them doing all kinds of bad things.

Turn now to a weak version of rationalism about moral motivation. According to a weak version, moral belief would entail having some motivation to the degree that one is practically rational. Consider now Rm_w:

> Rm_w: Belief that Φ-ing is morally right in circumstances C entails that one has some motivation to Φ in C in so far as they are practically rational.

Like R_w, Rm_w predicts that we will be conflicted when we take ourselves to have good reasons both for and against Φ-ing. Rm_w predicts, more precisely, that one will be conflicted in cases where moral reasons weigh against nonmoral reasons, but don't outweigh them. According to a weak version of rationalism about moral motivation, it's not surprising that Judah wrestles with whether to have his mistress killed. In so far as he is practically rational, he will have some motivation equivalent to the force of reason he has to kill or not to kill her. Since killing his mistress is a serious moral infraction, he will have a strong motivation not to kill her. But since killing also weighs heavily in favor of his long-term best interest, he will have a strong motivation to kill her. Ultimately, killing her is best to do all things considered. So his motivation to kill her will outweigh his motivation not to kill her in so far as he is practically rational.

But here's a problematic case for Rm_w. Suppose that Donny decides he would like to take candy from a small child. This, he can reasonably suspect, will cause the child to become very

upset. It would be morally wrong to do such a thing, he admits. Nonetheless the candy will give him some pleasure and that's really all he cares about. Donny is a sociopath. He does not care that he would hurt the child by taking her candy. He's also quite clever. He has figured out a way to take the candy without getting caught. He stands to gain slightly from doing something wrong. Is Donny irrational?

He is, in a sense. Donny is not acting with regard to what he has most reason to do. The small pleasure he stands to gain from taking candy from the child does not outweigh the harm that he will bring to the child. The moral implications of Φ-ing, in this case, outweigh the self-interested ones. According to Rm_w, Donny should have at least some motivation to resist taking the candy. But Donny is not in the slightest concerned about the child. He is not motivated at all to do what he has most reason to do.

However, though Donny is practically irrational in this instance, his disposition to ignore the concerns of others is quite rational. It would be rational to have this disposition, that is, if Donny had certain qualities. Suppose that Donny is a good moral disposition faker. He easily convinces people that he is genuinely concerned with the welfare of others. He expresses what seems to be sympathy when he encounters another's distress. He expresses what seems to be guilt over past moral transgressions he openly admits to. He becomes apparently outraged at injustice, indignant at slights, etc. But all of these attitudes are faked. He doesn't really feel indignance, shame, remorse, or any other moral sentiment. He just knows how other people express these attitudes and under what conditions they express these attitudes. When Donny does something wrong he doesn't really feel guilty about it. He just knows how to look like he feels guilty about it. Suppose further that Donny is very clever. He knows that taking advantage

of others is risky and requires great caution. He rarely seizes an opportunity to do so. But when he does he makes sure that he has a considerable likelihood of success.

Suppose further that Donny lives in a society like this one, where many of its members are disposed to do the right thing. This gives Donny ample opportunity to free ride. There's not too many people like Donny. Too many free riders spoil the advantages of being disposed to act morally. But if there are only a few people out there who regularly take advantage of others, then it remains in the best interest of most to act in ways that sometimes override what is in one's own best interest. Donny is in the minority. But this allows Donny the opportunity to take advantage when he thinks he can get away with it. Since he's also very clever, he usually does.

What would Donny's fully rational self think of Donny's disposition? Donny is disposed to act in ways that are best for himself. This is not the disposition of a moral person but it would be worse for Donny if he was disposed to do the right thing. Again, Donny is not like most of us. Given his situation among morally disposed people and his cleverness, he can afford to be insensitive to moral considerations. Donny's disposition is rational to have. His fully rational self would be glad. Thus, the disposition that it is rational to have may involve acting at times against what one has most reason to do. It can be rational, in other words, to be practically irrational on occasion.[85] In Donny's case, it is rational for him to have no motivation whatsoever to do the right thing, even though doing the right thing will be what he has most reason to do on some occasions. In cases like this, Donny's fully rational self would not advise him to obtain the motivation to do the right thing, for that would entail disposing him in ways that are worse for

85 This is a different kind of rational irrationality than what Parfit describes. The form he made famous is exemplified in most of us. For many of us, it is in our long-term best interest to act in ways that run counter to what is in our short term best interest. On the self interest theory of rationality, it is rational for us to be irrational on some or many occasions. See Parfit (1984: Section 5).

him. In order to maintain his self-serving disposition, Donny will lack motivation to do what is best, all things considered.

One might try to save Rm_w by arguing that Donny doesn't actually believe that Φ-ing is wrong to do and so his lack of motivation does not speak to any possible disconnect between belief and motivation. Like the way that Donny doesn't feel moral sentiment—that is, only in an inverted commas sense—Donny only believes that Φ-ing is wrong in an inverted commas sense. He doesn't believe that taking candy from a small child is wrong, but simply imitates what other people would say or think in this situation.

Certainly it could turn out to be the case that Donny believes only in an inverted commas sense. If, for instance, he didn't grasp how considerations that speak to his own long term benefit are reasons for him to act now, we might wonder if he understood how reasons that speak to another's concern could bear on how he is to act.[86] And there are sociopaths who fail to understand how anything could be a reason to act other than what is in their immediate, self-interested purview (Kennett 2002). But Donny, we may suppose, is not like these people. He understands all too well how his actions benefit himself in the long run. This, I think, is a mark of his cleverness. Further, that such people do exist seems plausible. There are probably more people like Donny in the world than many of us would like to admit.

Importantly, rationalism does not fail on an explanatory level. Recall that the best internalist position is one that restricts moral belief to certain kinds of agents, so-called 'normal' people. Moral belief is motivating in those who do not suffer some sort of psychological malady, such as sociopathy, or from undue stress. Rm_w would explain this phenomenon. Defeaters of Rm_w are people like Donny, sociopaths. Were one disposed to feel guilty for doing

86 Thus, Nagel argues that prudential reasons that bear on our future selves are similar to moral reasons (1971).

what he believed was wrong, the thought of doing wrong would harbor an emotional cost and provide some reason not to Φ. But Donny is not so disposed and so need not be concerned with this cost. Rm_w doesn't fail because it cannot explain moral motivation in 'normal' people. It fails because it doesn't apply equally to all people. Believing that Φ is right does not entail believing that one's fully rational self would advise one to be motivated, even to some degree, to Φ. In some cases, it is best to be disposed not to be motivated at all to do what is best.

3.5 Sentimentalism and Motivation

In this section I develop a sentimentalist account of moral motivation. The account here contrasts with an externalist model of moral motivation which explains the motivation-tracking phenomenon via a second order desire to do the right thing. On a sentimentalist reading, the content of moral judgment bears directly on what we are motivated to do. Sentimentalism turns moral conviction into an endorsement for feeling guilty. The normative force of the endorsement provides a rational constraint on what to feel, which in turn bears on motivation. A warrant constraint is narrower than the all-things-considered parameter of what is practically rational to do. It issues a reason to feel guilty, not a recommendation to do what one has most reason to do. But an endorsement to feel what is fitting is effective in producing the feeling. And feelings are effective in providing motivation.

An emotional account enjoys a structural advantage over a classical internalism. A classical internalism works from the most basic intuition of internalism: that reasons themselves are motivating actors (Nagel 1970, McDowell 1978). The suggestion is that normative or

justifying reasons can somehow move us. One way to carry out this suggestion is to posit a composite belief-desire state capable both of serving as a reason to Φ and a motivation to Φ, sometimes called a 'besire.' A besire is purportedly able to accommodate both aspects of belief and desire, but the notion is problematic. The difficulty lies in the different 'directions of fit' that beliefs and desires have, a phrase meant to describe the different ways that these attitudes represent their objects. Consider Smith's words on the subject:

> the difference between beliefs and desires in terms of direction of fit can be seen to amount to a difference in the functional roles of belief and desire. Very roughly, and simplifying somewhat, it amounts, *inter alia*, to a difference in the counterfactual dependence of a belief that p and a desire that p on a perception with the content that not p: a belief that p tends to go out of existence in the presence of a perception with the content that not p, whereas a desire that p tends to endure, disposing the subject in that state to bring it about that p.... Attributions of beliefs and desires require that different kinds of counterfactuals are true of the subject to whom they are attributed. (1994:115)

The belief that *p* is an attitude directed toward the world that represents some state of affairs. The desire that *p*, by contrast, is an attitude about how one wants the world to be. If I believe that *p*, I believe that the world is such that *p*, while if I desire that *p*, I want the world to be such that *p*. Accordingly, beliefs and desires are responsive to evidence of whether *p* in different ways. If I discover that *not-p*, for instance, my belief that *p* should change. But my desire may not. In fact, it's likely that I desire that *p* in part because the world is *not-p*.

Taken literally the notion of a besire seems incoherent. As Smith puts it, "a state with both directions of fit would therefore have to be such that, both, in the presence of such a perception it tends to go out of existence, and, in the presence of such a perception, it tends to endure, leading the subject who has it to bring it about that p" (1994:118). And this really doesn't make sense. The disparity in their counterfactual dependence on evidence against *p* renders the belief-desire composite as a state that has the tendency to both disappear and endure in light of that evidence. Further, this is somehow supposed to explain why we have the tendency to bring it about that *p*.

The defining characteristic of an internalism about motivation is that being disposed to Φ has some relation specifically to the content of the belief that Φ-ing is what one ought to do, morally speaking. But motivations and moral beliefs can come apart modally while still being internally related to each other. One way to account for this is to argue that, as Smith does, there is a rational connection between them. Believing that Φ would be morally right to do might be analyzed as believing that one should rationally desire to Φ. Belief and motivation would thus typically coincide in so far as one is practically rational. But, as I argued in the section above, the analysis doesn't altogether work, since it is not always rational to desire to do what one believes one morally ought to do.

Another way to achieve internal cohesion between moral belief and motivation is through rational sentimentalism. According to a rational version, moral belief is analyzed as the belief that guilt would be fitting. Moral belief thus invokes standards of warrant. Acceptance of these standards weigh on how we feel, which in turn bear on what we are motivated to do. The top-

down influence of moral belief on motivation travels from the standards we bring to bear on how we are to feel, and then from feeling to motivation.

The view suggests a rational influence of our beliefs over our feelings. Yet the kind of rationality involved is not necessarily practical. The concern over what to feel is not one which derives from a desire to do what one has most reason to do, or to feel what one has most reason to feel. Rather, the concern is over what we are warranted in feeling, or what feeling is appropriate given the circumstances. Call this kind of rationality 'emotional rationality.' One is emotionally rational to the extent that they feel what they think is fitting—if they respond to standards they place on themselves to feel what is appropriate.

There are, of course, various reasons to feel F, and it's rational to be responsive only to the best ones. But while other reasons may outweigh fittingness ones, only fittingness ones are of the right kind (Rabinowicz and Ronnow-Rasmussen, 2004). They speak to the logic of an emotion, as it were. Thus, there's a kind of oddity involved in an emotional response succored by reasons that are not the right ones. Consider the words of D'Arms on this subject:

> It's in the nature of these experiences to present themselves as sensitivities to something outside them. And what they present themselves as sensitivities to is a fairly restricted feature of the situation: a socially significant personal inadequacy, or a threat to one's safety, for instance. A little introspection makes it obvious, I think, that feelings of shame, fear, and so on just aren't about the advisability, or the moral permissibility, of feeling precisely that way. They are about a feature of the circumstance in virtue of which this is a fitting way to respond. In fearing the wild animal, I am struck by the

> things that make it fearsome (its size and ferocity, say), not by things that might make it wise or virtuous to be afraid. It may be neither wise nor virtuous to be afraid—if for instance, I'm the only thing between the wolves and the children.... Assessments of fittingness are attempts to make sense of or criticize our emotions using standards that speak to the distinctive concerns we take them to embody (2005:13).

Other kinds of reasons may justify feeling an emotion. There may be prudential reasons to feel F, and though these don't appear in the content of F they could be very good reasons. But as the remarks above suggest, invoking standards of warrant typically results in feeling the prescribed response. The thought that demands for responses along these lines appeal to justifications of the right kind is reinforced by the fact that it is odd for one to feel an emotion for reasons that don't speak to its fit.

Further, there is likely a prudential explanation for why fittingness reasons are often more effective than prudential reasons. Consider Frank's speculations about anger (1988). There are countless examples of people who act irrationally out of anger. There are those who do so fully aware that their actions are irrational. But if their anger is to lead them to irrational behavior, why indulge it? Wouldn't it be better to take measures to quell one's anger and avoid costly behavior? Though one should do something like this if one were to act in one's best interest, appealing to this kind of consideration is likely to do little good. Suppose, for example, that Harry is predisposed to become angry when wronged. Further, suppose that he gives signs to this effect. Perhaps he becomes gruff at a misunderstanding, or regularly complains loudly when someone cuts him off on the highway; he may even have a reputation for being someone with a

'short fuse.' Harry's disposition to anger is likely good for him in the long run. It would probably ward off anyone who knew that Harry would retaliate despite the considerable cost he might take on. Thus, it may be in Harry's best interest to be disposed to anger despite knowing that it is sometimes very irrational for him to be angry. And if anger wasn't resistant to strategic reasons for not feeling it, it couldn't play the role it does in deterring others from taking advantage of him.

An analogous tale can be told for guilt. Recall an earlier example of two people who could enter into a profitable venture together. Say that each will have the opportunity, at some later time, to cheat without the other noticing. Regardless of what the other does, it's in the best interest of each to cheat. However, if both cheat, both do worse than if neither cheats. But this means we have to act in ways that we know will later be contrary to our best interest. One way to solve this problem is by being disposed to keep one's promises. Guilt could serve in such a disposition. If each were disposed to feel guilty for going against their word, then the other could trust that they will not do so. But this means that one's guilt must not be sensitive to self-interested reasons for feeling it, since it will not be in one's best interest to feel guilty for cheating. Rather, guilt should be a sensitivity to considerations that bear on it's fit.

One who does not have the appropriate response is criticizable in the sense that he does not have what response it makes sense to have. We can say that one who does not feel what he thinks is appropriate does not feel what he should. Again, however, he may not be practically irrational. He does not feel, however, what would be justified by those reasons that speak to an emotion's fit. He does not feel what is rational to feel in this sense.

The feelings elicited by the rational constraint of moral belief bears on motivation. Moral beliefs are typically forward-looking, practical concerns about what to do. As such, guilt's role in moral belief is not direct. In believing that Φ would be wrong to do we do not feel guilty. Rather, we judge that it would be fitting to feel guilty were we to Φ. Thus, it is not guilt *per se* that keeps us from Φ-ing in such cases, but the prospects of feeling guilty that are based on fittingness reason we take ourselves to have to feel guilty.

These prospects only exist in agents who are disposed to feel guilty for doing what they believe is wrong. Those who are not disposed to feel guilty, such as sociopaths or listless people, will not feel guilty for doing what they believe is wrong, and so will not be motivated to resist acting immorally. The analysis thus applies only to 'normal' people, where 'normal' means those who are not sociopathic or listless. Describing the matter in this way is anticipated by a functional account of guilt. On this view, recall, guilt has the function of detecting what the agent believes to be instances of moral transgression. On this view, guilt can be said to misrepresent its object if it fails to coincide with our moral beliefs. This provides a normative basis for describing those who are disposed to feel guilty for doing what they believe is wrong as 'normal.' 'Abnormal' people, on the other hand, are those in whom guilt is not functioning as it should.

Importantly, what counts as 'normal' is not whatever social norms would proscribe. One can dissent from popular opinion and still be 'normal' for all that. Rather, 'normality' is an agent-relative parameter. It consists in whatever one believes to be right to do and the corresponding tendency to feel guilty for not doing it. One may therefore be radical in her dissent from cultural

norms by challenging the status quo, demanding a different set of standards for what counts as right or wrong to do, and still be 'normal.'

We can also distinguish between those things that are actually wrong to do (if any), and those things we are disposed to feel guilty for doing. A 'normal' person is disposed to feel guilty for doing what she believes is wrong, a disposition we can identify with the function of guilt. But the function of guilt is tied to what we believe, not necessarily to the facts. And where beliefs do not line up with the facts, neither do dispositions to feel guilty line up with the facts. There is a difference, in other words, between an emotion's fittingness and its function.

Return again to Rabinowicz and Ronnow-Rasmussen's (2004) solution to the wrong kinds of reasons problem. According to their view, the right kinds of reasons for having a particular response have a dual role. A response is fitting, first, if its object has the property that appears in its content. Second, the property of an object that appears in the content of a response must justify having that response. If Φ-ing is morally wrong, then, it satisfies the dual role of reasons that would justify feeling guilt (it's fitting). Φ-ing's being wrong to do both appears in the content of guilt and justifies feeling it. It's the kind of reason that speaks to the fittingness of guilt. But suppose that I think that slavery is a morally acceptable practice, something that, presumably, I'm wrong about. If this were the case, and I am 'normal,' then I will not be disposed to feel guilty for something that I should, and so my indifference to the practice is unfitting. Nonetheless, it's consistent with what I believe is wrong to do. Since guilt has the function of detecting what I believe is wrong, then, in as much as my acceptance of slavery is consistent with what I believe, it does what it is supposed to do. I should, however, be disposed

feel guilty for endorsing the practice of slavery. There are reasons of the right kind that call for it.

Thus, fittingness and function may never go together, or conversely, they may always go together. Suppose that, instead of one set of universally binding moral facts, there is no fact of the matter about whether Φ-ing is morally wrong to do. If this were the case, then guilt would never be fitting. Yet it could function well enough. It could track the norms one uses to govern his own behavior. It's just that, every time he felt guilty, he would be under the sway of a response that is not appropriate. There is no object that has the property moral wrongness and no reasons of the right kind for guilt. On the other hand, norm and fact may always go together. It may be the case, for instance, that the moral facts are individually relative. If this were true, whatever norms one adopted to govern her responses would constitute the moral facts. So long as guilt tracked her norms of acting rightly there could never be an instance in which it is unfitting. Her norms constitute the facts. Every time her disposition to feel guilty lines up with the norms, it lines up with the facts.

In addition to being disposed to feel guilty for doing what one believes is wrong one must also be able to appreciate the emotional cost of Φ-ing. The prospect of feeling guilty is not so good. This cost should weigh into what the agent has reason to do in so far as she is rational. Accordingly, one could be irrational in this sense. We might imagine someone who would be willing to say, 'I know I would feel guilty for stealing a cookie from the cookie jar, but so what? I don't care about future emotional discomfort. All that matters is what I would gain now by having the cookie.' This ever-present minded person would be disposed to feel guilty for doing what she believes is wrong and yet fail to be motivated to resist doing it. If this character is

coherent, moral motivation requires both the disposition to feel guilty for acting against one's moral judgment and a sensitivity to the emotional cost of feeling guilty.

Thus, the motivational connection to believing morally wrong consists in what an emotionally rational person would do were she sensitive to the emotional cost of acting against her beliefs. Consider now E:

> E: Belief that Φ-ing is morally wrong entails that one will be motivated to resist Φ-ing in so far as she is emotionally rational and sensitive to the emotional cost of Φ-ing.

Like other forms of internalism, E can be given a strong or weak reading. The strong version would have it that believing morally wrong entails having an overriding motivation to resist Φ-ing in so far as she is emotionally rational and sensitive to the emotional cost of Φ-ing. This view, however, is unable to account for cases in which other reasons for or against feeling guilty outweigh the fittingness rationale. Emotional rationality pertains to one's disposition to feel guilty. Thus, E is a claim that has implications not just for what one is disposed to feel, but also for how that disposition in addition to a sensitivity to emotional cost bears on what one is motivated to do. On a strong reading of E, the disposition to feel guilty and corresponding sensitivity results in an overriding motivation to resist Φ-ing. But one can be emotionally rational and sensitive to cost without having such a motivation. Judah believes that killing his mistress is morally wrong and that he should feel guilty for doing it. In so far as he is emotionally rational he will be disposed to this response. Nonetheless, if his mistress is threatening to ruin his life, then it would be best for Judah to take on the extreme pangs of guilt

and go through with the deed anyway. Removing his mistress would alleviate the threat only if he does not turn himself in for having done so. By resisting the impulse to make reparation for his wrongdoing, Judah produces the best outcome for himself. He has fittingness reason to feel guilty and is accordingly disposed and sensitive to its cost, but does not have, and rationally should not have, an overriding motivation not to Φ.

A weak version can accommodate the play between fittingness reasons for guilt and prudential reason to do what one believes guilt would be a fitting response to. On a weak view, one will only have some motivation to resist Φ-ing in so far as one is emotionally rational and sensitive to cost. Thus, so long as one is disposed to feel guilty for doing what one believes is wrong and bothered by the prospect, other reasons in favor of Φ-ing could outweigh the endorsement of warrant. Further, since fittingness reasons are to provide one with some motivation under these conditions, the agent will be conflicted in these cases about what to do. Though Judah may not have an overriding motivation to resist killing his mistress, he is bothered by the thought of doing so. His motivation to resist Φ-ing derives from considerations that bear on whether he should feel guilty and the prospect of taking on this emotional burden. But the cost of letting her live are too high (unfortunately), and so is conflicted about what to do.

E is thus most plausibly construed as a weak thesis about motivation. Consider finally E_w:

> E_w: Belief that Φ-ing is morally wrong to do entails having some motivation to resist Φ-ing in so far as one is emotionally rational and sensitive to the emotional cost of Φ-ing.

E_w provides the link between moral belief and motivation in 'normal' people. It eschews outliers embodied by sociopathy and listlessness from the analysis, and provides a normative basis for criticizing those who in whom the disposition to feel guilty does not match up to their judgment about the fittingness of guilt. As well, it is not so strong as to make the connection one between what we believe is right and motivations that we would be best advised to have. E_w is, in other words, a form of rationalism, but not one that overstates the connection our moral beliefs have to what we are motivated to do.

As an internalism about motivation, E_w allows for the sensitivity to grounds for motivation that externalism lacks. Where fittingness reasons are taken to bear on what to feel, we take ourselves to have grounds for being motivated to resist Φ-ing. But where these grounds are perceived to disappear or absent, we take ourselves to have no reason to be motivated to resist Φ-ing. Thus, E_w explains the motivation-tracking phenomenon as it occurs in both directions, both as a tendency for motivation to change in response to believing morally wrong and in response to believing morally permissible. The claim meets the desiderata for an account of moral motivation, and stands as a credit to sentimentalism's ability to solve the problem of moral motivation.

References

Ayer, A.J. 1936: *Language, Truth and Logic.* Gollancz.

Arendt, Hannah 1977: *Eichmann in Jerusalem: A Report on the Banality of Evil.* Penguin.

Blackburn, Simon 1998: *Ruling Passions: A Theory of Practical Reasoning.* Oxford University Press.

1993: *Essays in Quasi-Realism.* Oxford University Press.

1987: 'How to Be an Ethical Antirealist' in Peter A. French, Theodore E. Uehling, Jr. and Howard K. Wettstein, eds, 1987: *Midwest Studies in Philosophy Volume XII: Realism and Anti-realism*. University of Notre Dame Press. 361-75.

1973: 'Errors and the Phenomenology of Value' reprinted in Blackburn 1993.

1980: 'Opinions and Chances' in D.H. Mellor, ed., *Prospects for Pragmatism.* Cambridge University Press.

Brandt, Richard 1946: 'Moral Valuation', *Ethics.* 106-21.

Brentano, Franz 1889/1969: *The Origin of Knowledge of Right and Wrong*, trans. Roderick Chisolm. Routledge and Kegan Paul.

Brink, David O., 1997: 'Moral Motivation', *Ethics.* 4-32.

1989: *Moral Realism and the Foundation of Ethics.* Cambridge University Press.

Cleckley, Harvey, 1964: *The Mask of Sanity.* 4th ed., St. Louis: Moseby.

Copp, David 1997: 'Belief, Reason, and Motivation: Michael Smith's *The Moral Problem*',

Ethics. 33-54.

Crisp, Roger 2000: review of *Value and What Follows*, by Joel Kupperman. *Philosophy* 75: 458-62.

Damasio, Antonio 1994: *Descartes' Error.* G. P. Putnam's Sons.

D'Arms, Justin 2006: 'Two Arguments for Sentimentalism', *Philosophical Issues,* 1-21.

and Jacobson, Daniel 2005: 'Anthropocentric Constraints On Human Value', forthcoming in *Oxford Studies in Metaethics 1.*

2003: 'The Significance of Recalcitrant Emotions (Or Anti-QuasiJudgmentalism)' in *Philosophy and the Emotions,* A. Hatzimoysis, ed., (Cambridge:Cambridge University Press).

2000a: 'Sentiment and Value', *Ethics.* 722-48.

2000b: 'The Moralistic Fallacy: On the Appropriateness of Emotions', *Philosophy and Phenomenological Research.* 65-90.

1994: 'Expressivism, Morality and the Emotions', *Ethics.* 739-63.

Darwall, Stephen 1997: 'Reasons, Motives, and the Demands of Morality: An Introduction', in Allan Gibbard, Peter Railton, eds., *Moral Discourse and Practice: Some Philosophical Approaches*. 305-312.

Davidson, Richard J.1994: *The Nature of Emotion: Fundamental Questions.* Oxford University Press.

1999: 'Basic Emotions,' in T. Dalgleish & T. Power, eds., *The Handbook of Emotion and Cognition.* Wiley.

Dreier, James 2000: 'Dispositions and Fetishes: Externalist Models of Moral Motivation',

Philosophy and Phenomenological Research. 619-38.

1990: 'Internalism and Speaker Relativism', *Ethics.* 6-26.

Dretske, Fred 1995: *Naturalizing the Mind.* MIT Press.

Ekman, Paul and Friesen, Wallace 1971: Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology. 17.* 124-129.

Ewing, A.C. 1939: "A Suggested Non-Naturalistic Analysis of Good." *Mind* 48:1-22.

1947: *The Definition of Good*. Mackmillian Company.

Falk, W. D. 1948: '"Ought" and Motivation', *Proceedings of the Aristotelian Society.* 111-38.

Frank, Robert H. 1988: *Passions Within Reason: The Strategic Role of the Emotions.* W. W. Norton & Company.

Gauthier, David 1985: *Morals by Agreement.* Oxford University Press.

Gibbard, Allan 1990: *Wise Choices, Apt Feelings*. Clarendon Press.

Greenspan, Patricia 1988: *Emotions and Reason*. Routledge.

Griffiths, Paul E. 1997: *What Emotions Really Are.* Chicago: University of Chicago Press.

Hare, R. M. 1952: *The Language of Morals.* Oxford University Press.

Harman, Gilbert 1977: *The Nature of Morality.* Oxford University Press.

Hauser, M. D. 2002: *Wild Minds: What Animals Really Think.* Penguin.

Hieronymi, Pamela 2005: 'The Wrong Kind of Reason', *The Journal of Philosophy.* 437-457.

Hume, David 1965: 'Of the Standard of Taste', in John W. Lenz, ed., *Of the Standard of Taste and Other Essays*. The Bobs-Merrill Company. 3-24.

1739/1978: P.H. Nidditch, ed., *A Treatise of Human Nature.* Clarendon Press.

Kennett, Jeanette 2002: 'Autism, Empathy and Moral Agency', *The Philosophical*

Quarterly. 340-357.

Kenny, Anthony 1963: *Action, Emotion and Will.* Routledge and Kegan Paul.

Korsgaard, Christine 1986: 'Skepticism about Practical Reason', *Journal of Philosophy.* 5-25.

Lazarus, R. S. 1991: *Emotion and Adaptation.* Oxford University Press.

Lewis 2000: 'The Emergence of Human Emotions,' In M. Lewis & J. M. Haviland-Jones, eds., *Handbook of Emotions.* 2nd ed., 265-280. Guilford Press.

Lillehammer, Hallvard 1997: 'Smith on Moral Fetishism', *Analysis.* 187-95.

Mackie, J. L. 1977: *Ethics: Inventing Right and Wrong.* Penguin. Ethica. Cambridge University Press.

McDowell, John 1987: 'Projection and Truth in Ethics', The Lindley Lecture, The University of Kansas.

Mele, A. 1996: 'Internalist Moral Cognitivism and Listlessness', *Ethics.* 727-53.

Moore, G. E. 1903: *Principia Ethica.* Cambridge University Press.

Nussbaum, Martha 1994: *The Therapy of Desire: Theory and Practice in Hellenistic Ethics.* Princeton University Press.

Olson, Jonas 2004: 'Buck-Passing and the Wrong Kinds of Reasons', *The Philosophical Quarterly.* Vol 54. No. 215. 295-300.

Parfit, Derek 2001: 'Rationality and Reasons,' in *Exploring Practical Philosophy: From Action to Values*, Ed. Dan Egonsson, Bjorn Petersson, Jonas Josefsson, and Toni Ronnow-Rasmussen. Ashgate. pp. 17-41.

1984: *Reasons and Persons.* Oxford University Press.

Prinz, J. 2004a: *Gut Reactions.* Oxford University Press.

2004b: *The Emotional Construction of Morals.* Manuscript.

Rabinowicz, Wlodek and Rønnow-Rasmussen 2004: 'The Strike of the Demon: On Fitting Pro-Attitudes and Value', *Ethics.* 391-423.

Railton, Peter 1997: 'Aaesthetic Value, Moral Value, and the Ambitions of Naturalism', reprinted in Railton 2003.

2003: *Facts, Values and Norms.* Cambridge University Press.

Rawls, John 1971: *A Theory of Justice.* Harvard University Press.

Roberts, Robert 1988: 'What an Emotion is: A Sketch', *Philosophical Review.* 183-209.

Rorty, Amelie 1980: Introduction. In *Explaining Emotions*. University of California Press.

Schmidtz, David 1994: 'Choosing Ends', *Ethics*. 226-51.

Scanlon, T. M. 1998: *What We Owe to Each Other.* Harvard University Press.

Smith, Michael 1997: 'In Defense of *The Moral Problem*: A Reply to Brink, Copp, and Sayre-McCord', *Ethics*, 84-119

1995: 'Internal Reasons', *Philosophy and Phenomenological Research*. 109-131.

1994: The Moral Problem. Blackwell.

Solomon, Robert 1976: *The Passions.* Anchor/Doubleday.

Stevenson, C. L. 1937: 'The Emotive Meaning of Ethical Terms', *Mind.* 14-31.

Stocker, Michael 1979: 'Desiring the Bad: An Essay in Moral Psychology', *Journal of Philosophy.* 738-53.

Sturgeon, Nicholas 1986: 'What Difference Does it Make Whether Moral Realism is True?', *Southern Journal of Philosophy.* 115-141.

Tooby, J., & Cosmides, L. 1990: 'The past explains the present: Emotion adaptations and the

structure of ancestral environment. *Ethology and Sociobiology.* 375-424.

Wiggins, David 1987: 'A Sensibile Subjectivism' in his *Needs, Values Truth.* Basil Blackwell. 184-214.

Williams, Bernard 1981: *Moral Luck*. Cambridge University Press.

1980: 'Internal and External Reasons', reprinted in Williams 1981. 101-13.

1976: 'Persons, Character and Morality' reprinted in Williams 1981. 1-19.